PHM-Based Modeling for Cyberattack Classifier Performance

Priscila Silva

Department of Electrical and Computer Engineering, University of Massachusetts Dartmouth, Dartmouth, MA, 02747, USA

psilva4@umassd.edu

ABSTRACT

This research implements Prognostics and Health Management (PHM) using multiple linear regression and multivariate time series models to monitor and predict when the performance of a Machine Learning-based cyberattack classifier might degrade to an unacceptable level, enabling preemptive maintenance strategies.

1. INTRODUCTION

A Network Intrusion Detection System (NIDS) (Liao et al., 2013) analyzes network traffic to detect suspicious patterns or anomalies, alerting administrators to potential threats. Deep neural networks (DNNs) (Ahmad et al., 2020) are Machine Learning (ML) techniques often employed in NIDS for their ability to accurately classify cyberattacks (Lewis, 2002) such as malware infections and denial-of-service attacks. Although DNNs perform well in identifying known attacks (Baye et al., 2023), their resilience against unknown activities is less studied.

Problem to be addressed: Past research (Javaid et al., 2016; Sharma et al., 2019; Wu et al., 2020; Narayana Rao et al., 2021; Lo et al., 2022) has explored various techniques to enhance the robustness of DNNs. However, these techniques often rely on specific benchmark datasets, leading to incomplete representations of real-world network settings, and require long training times, posing challenges in promptly and cost-effectively identifying cyberattacks. Furthermore, there is a lack of quantitative assessment regarding the reliability of these techniques' predictions over time. Without predictive models that can monitor classifiers in real-time and forecast future performance, anticipating new threats and adapting NIDS performance strategies may be challenge.

Expected novel contributions: This research implements Prognostics and Health Management (PHM) techniques, including multiple linear regression (Kleinbaum, Kupper, Nizam, & Rosenberg, 1999) and multivariate time series models (Brandt & Williams, 2007) approaches, to monitor and predict the performance of DNNs based on the proximity of incoming real-time cyberattacks to known classes. Anticipating the performance of classifiers might streamline the testing process with new datasets, reducing evaluation time in the face of unknowns. Additionally, it aids in monitoring and assurance for NIDS, empowering professionals to gauge NIDS performance trends, proactively address potential performance degradation, and identify optimal maintenance strategies to sustain performance.

Proposed research plan: The proposed research, initiated in Fall 2022 and scheduled for completion by Spring 2026, encompasses a comprehensive plan within the PhD program. The main activities include: (i) model changes in the performance of cyberattack classifiers with different PHM techniques (2022); (ii) enhance prediction capabilities through improved parameter estimation techniques (2023); (iii) formulate optimization problems to identify resilience requirements such as maintenance schedules (2024); (iv) retrain classifiers based on identified resilience requirements to rapidly and efficiently restore classifier performance after degradations (2025); and (v) submit manuscript and defense dissertation (2026). In sum, this PhD dissertation proposes to ensure the continuous and reliable operation of ML-based cyberattack classifiers using PHM techniques, which is crucial for enhancing system resilience by identifying the best timing for interventions to effectively mitigate risks and recover from adversarial attacks.

2. CYBERATTACK CLASSIFIERS

A deep neural network (DNN) consists of interconnected layers of neurons governed by mathematical functions. In NIDS, DNNs receive NIDS benchmark payload data - packet sections transmitting network information often hiding malware - in the input layer, pass it through hidden layers to extract features, and make predictions at the output layer. Training DNNs to learn attack patterns involves refining hyperparameters for optimal performance using evaluation metrics such as the F1-Score. The F1-Score is a reliable measure of NIDS classifier performance since it measures the overall model balance between identifying true cyberattacks among

all instances classified as attacks and detecting all true cyberattacks. In live, when new instances are fed into a trained classifier, it classifies whether each packet is malicious or benign based on the patterns it has learned during training. After classification, samples are clustered based on predicted labels, with distance metrics quantifying data point similarities or differences within clusters. Typical distance metrics encompass the *Euclidean distance*, which gauges the straight distance between a new sample and the closest known cluster's mean; *cosine similarity*, which calculates the cosine of the angle between the new sample and the nearest known cluster's mean; and *Intra-cluster spread*, which evaluates the dispersion of samples within a cluster, signifying the extent to which the cluster expands upon receiving new samples.

3. PROGNOSTICS AND HEALTH MANAGEMENT OF CLASSIFIERS

Prognostics and Health Management (PHM) of cyberattack classifiers involves implementing techniques to monitor, assess, and sustain the performance of these classifiers over time. For example, Figure 1 illustrates the performance (P) of a cyberattack classifier trained on a dataset containing various types of cyberattacks. Initially, after training and testing the classifier, it achieves a F1-Score of $P = 0.9$, indicating a high performance at classifying cyberattacks. Over time, new network packets with patterns and characteristics different from the original training data begin to emerge. Consequently, the classifier's ability to accurately classify these new packets as benign or malicious starts to decline, reducing the F1-score from its initial high value of $P = 0.9$ to a lower value, such as $P = 0.6$. This lower value indicates a warning state where restorative actions are necessary in order to maintain the classifier's usability. At this point (t_{m_1}), with an explanatory diagnostic, maintenance techniques such as retraining the DNN including information learned from the new threats or updating the classifier parameters to improve its effectiveness could restore the classifier's performance, as illustrated by the dashed line. Otherwise, the system will continue to operate with low performance only during its remaining useful life (RUL) until it degrades to an unacceptable level, illustrated as $P = 0.5$, where the DNN is no longer reliable.

4. PREDICTIVE MODELS

Regression and time series models are suitable techniques for predicting the performance (P) of DNNs, defined as the F1-score. By forecasting the F1-score, these models can aid decision-makers in optimizing cyberattack detection while minimizing false alarms and missed detections, which is crucial for network security. Regression models predict the F1-score by analyzing its linear relationship with covariates representing the distance between new instances and known classes of cyberattacks. Techniques include multiple linear regression (MLR) and multiple linear regression

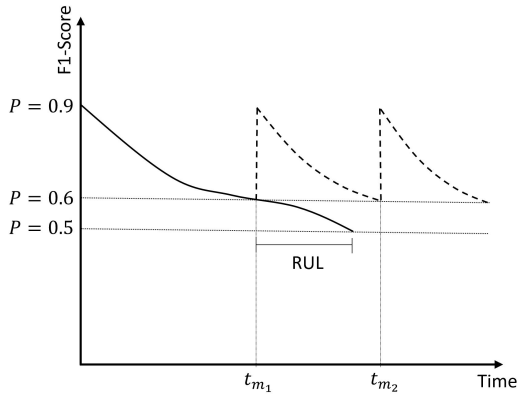

Figure 1. Performance analysis of cyberattack classifiers

with interaction (MLRI), both assuming normally distributed residuals for reliable conclusions and generalization. Multivariate time series models, such as multivariate vector autoregressive (MVAR) and multivariate vector auto-regressive moving average (MVARMA), forecast F1-score changes over time based on past F1-scores and distances between known and new samples. These models require data to be stationary, which can be confirmed through time sequence graphs.

Table 1 presents each predictive model discussed above to predict F1-score at time interval i. In each equation shown in the Table, β_0 is the average value of F1-score, X_j are the $j = 1, \ldots, m$ covariates representing the distances between new threats and the centroid of the known classes that the DNN was trained on, β_j is the covariates impact on F1-score, $\beta_{j(m+k)}$ represents the interaction between covariates X_j and X_k, β_k describes the F1-score relationship to intervals $(i - \ell) \leq (i-k) \leq (i-1))$, $\beta_{j(\ell+k)}$ is the impact of the distances of samples in the previous $(i - k)$ intervals, and θ_k is the coefficient associated with k times steps before the present time step of a sequential white noise process (ε), which are statistically independent, and normally distributed with zero mean and finite variance.

Model fitting is performed with least squares estimation (LSE) (Pham, 1999), which minimizes the discrepancy between the actual and predicted F1-score data to estimate all parameters present in the models shown in Table 1. Evaluating a model's performance on the given dataset involves computing various goodness-of-fit measures, including: (i) sum of squares error (SSE), (ii) Predictive mean square error (PMSE), (iii) adjusted coefficient of determination (r^2_{adj}), and (iv) confidence intervals to establish a range for estimated values based on a user-specified level of confidence.

5. PRELIMINARY RESULTS

The proposed modeling approaches are illustrated using experiments of a classifier pre-trained on 70% of randomly selected samples of 10 classes from the NIDS benchmark CIC-IDS2017 dataset (Rosay, Cheval, Carlier, & Leroux, 2022), which contains 1,410,255 samples under 15 different classes

Table 1. Regression and Time Series Models to Predict F1-score

Approach	Model	Equation
Regression	MLR	$\hat{P}(i) = \beta_0 + \sum_{j=1}^{m} \left(\beta_j X_j(i)\right)$
	MLRI	$\hat{P}(i) = \beta_0 + \sum_{j=1}^{m} \left(\beta_j X_j(i)\right) + \sum_{j=1}^{m} \left(\sum_{k=j+1}^{m} \left(\beta_{j(m+k)} X_j(i) X_k(i)\right)\right)$
Time Series	MVAR	$\hat{P}(i) = \beta_0 + \sum_{k=1}^{\ell} \left(\beta_k P(i-k)\right) + \sum_{j=1}^{m} \left(\sum_{k=1}^{\ell} \left(\beta_{j(\ell+k)} X_j(i-k)\right)\right)$
	MVARMA	$\hat{P}(i) = \beta_0 + \sum_{k=1}^{\ell} \left(\beta_k P(i-k)\right) + \sum_{j=1}^{m} \left(\sum_{k=1}^{\ell} \left(\beta_{j(\ell+k)} X_j(i-k)\right)\right) + \sum_{k=1}^{\ell} \left(\theta_k \varepsilon(i-k)\right)$

of cyberattacks. The remaining 30% of the data was used for validation and testing. To predict the F1-score of the pre-trained DNN using the models discussed in Section 4, a new dataset was collected through various experiments designed to assess the algorithm's performance across diverse scenarios. Each experiment involved testing the classifier on a fresh, balanced dataset containing instances from the 10 known classes and the remaining 5 unknown classes, which were not part of the training set, to gauge how well the classifier performed with previously unseen data. In each experiment, the F1-score was computed to evaluate the classification performance. Additionally, distance metrics discussed in Section 3 were calculated, representing the average distances from all new samples to their respective nearest known class. These metrics were recorded as input data for the predictive models, with the F1-score designated as the performance to be predicted (P), and the distance metrics as candidate covariates (X) of the model. By estimating the model parameters using the data gathered from these experiments, forecasting the future F1-score of a classifier involves computing the distance between new real-time instances and the known attack patterns stored in the historical data.

The model development follows these steps: (i) Create initial regression and time series models with all (no) covariates and correlated lag identified from autocorrelation function (ACF) and partial autocorrelation function (PACF). (ii) Use least squares estimation with 80% of data, compute goodness-of-fit measures, and validate models using the remaining 20% data for prediction accuracy assessment. (iii) Add or remove covariates using forward and backward stepwise procedures until minimizing $PRMSE$. (iv) Increase lags for F1-score and covariates following ACF and PACF, repeat steps (ii)-(iv) until no improvement in $PRMSE$. (v) Choose the model with the lowest $PRMSE$ for predicting F1-score.

Table 2 reports the covariates (X) included in each model, where $(X1)$ is the Euclidean distance, $(X2)$ is the cosine distance, and $(X3)$ is the intra-spread distance, as well as the number of homogenous number of lags (ℓ) of all features that minimized the $PRMSE$, indicating for example that MVAR model includes 3 lags of the previous performance P, as well as 3 lags of each covariate $X3$ and $X1$. Table 2 also shows the number of parameters (p) contained in each model, and

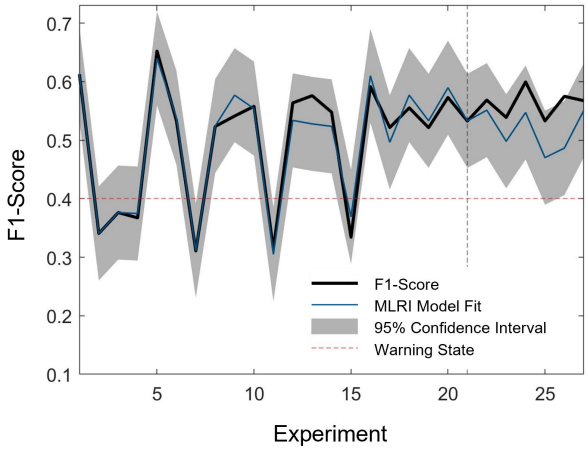

Figure 2. MLRI model fit and its 95% confidence interval

the associated goodness-of-fit values achieved, which are penalized for the number of parameters in the model. The best-performing model, highlighted in bold in Table 2, is the regression model with covariate interactions. This model outperformed others in all goodness-of-fit measures, indicating its superior ability to monitor and predict the F1-score trend effectively. While time series models offered a comprehensive framework and capture small changes in performance, simpler models like MLRI could capture essential data characteristics with fewer parameters, making them more practical and easier to interpret for future F1-score predictions.

To visualize the best model fit, Figure 2 shows the empirical data as well as the fitted MLRI model and its 95% confidence intervals (grey region), where the dashed vertical line at 80% of the data marks the end of model fitting and the start of predictions. Since all but 1 of the 27 data points fell within the confidence interval, the empirical coverage is 96.29%, which indicates that the MLRI model was able to monitor and predict both decreasing and increasing trends of the F1-score well, aiding in understanding new sample attributes. This result suggest that MLRI is suitable to identify maintenance strategies to address unknown cyberattacks. The dashed red line represents a hypothetical warning threshold, part of the ongoing research in this dissertation to pinpoint the best time for classifier maintenance.

Table 2. Validation of Model's Prediction

Model	Covariates	Lags (ℓ)	Parameters (p)	$RMSE$	$PRMSE$	r^2_{adj}
MLR	X3, X1, X2	0	4	0.0687	0.0685	0.7344
MLRI	X3, X1, X2	0	7	**0.0409**	**0.0488**	**0.9077**
MVAR	P, X3, X1	3	10	0.0441	0.0506	0.7829
MVARMA	P, X1, X2, X3	1	12	0.0911	0.1164	0.4115

6. CONCLUSION AND FUTURE DIRECTION

This research proposes to apply Prognostics and Health Management (PHM) methods using statistical models to monitor and predict the performance of cyberattack classifiers against new threats. Two regression models and two multivariate time series models were tested, aiming to accurately predict the F1-score based on real-time distances between new threats and known cyberattacks. Preliminary results indicate that while time series models can adjust to data fluctuations, simpler regression models can effectively capture data characteristics, potentially offering sufficient insights for F1-score predictions without added complexity. Specifically, multiple linear regression with interaction between covariates showed a high empirical coverage of 96.29%, suggesting accurate predictions of future observations. This approach may support NIDS researchers and practitioners to optimize classifier performance against specific threats by identifying and adjusting training methods or detection strategies. Future work will explore additional features in models to enhance F1-score predictions and integrate these predictions into decision-making for optimal classifier maintenance and replacement strategies against new cyber threats.

ACKNOWLEDGEMENT

This work was supported in part by the U.S. Military Academy under Cooperative Agreement No. W911NF-22-2-0160. The views and conclusions expressed in this paper are those of the authors and do not reflect the official policy or position of the U.S. Military Academy or U.S. Army.

REFERENCES

Ahmad, Z., Khan, A. S., Shiang, C. W., Abdullah, J., & Ahmad, F. (2020). Network intrusion detection system: A systematic study of machine learning and deep learning approaches. *Transactions on Emerging Telecommunications Technologies*, *32*, e4150.

Baye, G., Silva, P., Broggi, A., Fiondella, L., Bastian, N. D., & Kul, G. (2023). Performance analysis of deep-learning based open set recognition algorithms for network intrusion detection systems. In *Noms 2023-2023 ieee/ifip network operations and management symposium* (p. 1-6).

Brandt, P., & Williams, J. (2007). *Multiple time series models* (No. no. 148). SAGE Publications.

Javaid, A., Niyaz, Q., Sun, W., & Alam, M. (2016). A deep learning approach for network intrusion detection system. In *Proceedings of the 9th eai international conference on bio-inspired information and communications technologies (formerly bionetics)* (pp. 21–26).

Kleinbaum, D. G., Kupper, L. L., Nizam, A., & Rosenberg, E. S. (1999). *Applied regression analysis and other multivariable methods*. Cengage Learning.

Lewis, J. A. (2002). *Assessing the risks of cyber terrorism, cyber war and other cyber threats*. Center for Strategic & International Studies Washington, DC.

Liao, H.-J., Richard Lin, C.-H., Lin, Y.-C., & Tung, K.-Y. (2013). Intrusion detection system: A comprehensive review. *Journal of Network and Computer Applications*, *36*(1), 16-24.

Lo, W. W., Layeghy, S., Sarhan, M., Gallagher, M., & Portmann, M. (2022). E-graphsage: A graph neural network based intrusion detection system for iot. In *Noms 2022-2022 ieee/ifip network operations and management symposium* (p. 1-9).

Narayana Rao, K., Venkata Rao, K., & P.V.G.D., P. R. (2021). A hybrid intrusion detection system based on sparse autoencoder and deep neural network. *Computer Communications*, *180*, 77-88.

Pham, H. (1999). Software reliability. *John Wiley & Sons*.

Rosay, A., Cheval, E., Carlier, F., & Leroux, P. (2022). Network intrusion detection: A comprehensive analysis of cic-ids2017. In *8th international conference on information systems security and privacy* (pp. 25–36).

Sharma, J., Giri, C., Granmo, O.-C., & et al. (2019). Multi-layer intrusion detection system with extratrees feature selection, extreme learning machine ensemble, and softmax aggregation. *EURASIP Journal on Information Security*, *2019*(15). doi: 10.1186/s13635-019-0098-y

Wu, H., Sun, P., Liu, P., Li, Q., Liu, C., Lu, X., ... Chen, J. (2020). Dl-ids: Extracting features using cnn-lstm hybrid network for intrusion detection system. *Security and Communication Networks*, *2020*, 8890306.

BIOGRAPHY

Priscila Silva is a Ph.D. candidate in Electrical and Computer Engineering at the University of Massachusetts Dartmouth. She is an advisee of **Dr. Lance Fiondella** and **Dr. Gokhan Kul**, working in projects supported by the U.S. Military Academy and the U.S. Department of the Air Force.