IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Securing Health Sensor Streams with ML-Based Malware Classifiers

[1]GALI NAGAMMA,
Student in Dept. Of Master of Computer Applications, at Miracle Educational Society Group of Institutions
[2]A BHAVANI, Miracle Educational Society Group of Institutions
[3]Dr. INDU, Miracle Educational Society Group of Institutions
[1]nagammagali312@gmail.com

ABSTRACT:

The integration of health sensor technology within the Internet of Things (IoT) framework poses unprecedented risks with embedded malware. The problem of outdated traditional monitoring systems failing to detect mutated modern malware variants exists. Using health sensor data, this project incorporates a machine learning technique to identify hidden malware, utilizing sophisticated classification models: XGBoost, LightGBM, and Random Forest. The system classifies malware through bytes of static analysis. The features to the system are encoded to detect patterns in the code with malicious intent, and thus are trained through a learning algorithm to yield precise predictions. Effective models are created and tested against health datasets, resulting in new models that outperform traditional methods in both detection and classification tasks of malicious code in health systems.

Keywords: detecting malware, IoT, XGBoost

INTRODUCTION

Nowadays, the data collection and integration of health monitoring systems paired with the Internet of Things (IoT) has become very common, allowing data to be collected in real time and aiding in monitoring patients through sensors. The rise of such systems facilitates collection and transmission of patients' data, but also exposes such systems to malware, which can harm data integrity, infringe upon sensitive operation, or disrupt processes. Malware embedded in sensor data often employs sophisticated, evasive techniques to avoid detection by traditional security systems. This underscores the urgency of developing effective countermeasures to detect such malicious activities. This project aims to implement machine learning algorithms in the static code analysis of malware patterns within health sensor data. The system is capable of predicting malware presence through classification based on labeled datasets that contain benign and malware samples

at the byte and token levels. The objective is to create a malware detection system that is light weight, precise, and efficient in protecting health sensor ecosystems from cyber threats.

RELATED WORK

Other Studies in the area of malware detection and safeguarding sensor data have been done. A feature subset selection approach based on correlation for anomaly detection in IoT environments was presented by Su et al. (2019). focus on real-time responsiveness and relevance-driven data filtering for enhancement of learning model accuracy was the focus of their work. Yu et al. (2018) developed CBDLP, a data leakage prevention model. It constructs confidential and contextual term graphs to identify partially concealed term leaks in mobile devices. Sun et al. (2018) introduced a secure data-sharing framework for darknets that is selective in the construction of routing paths through controlled node hierarchies using hierarchical greedy embedding. Wang et al. (2018) proposed the use of attack path reconstruction as a timely mitigative approach through asynchronous response to RDP-based malware and provided ransomware detection through traced attack paths. Finally, Xiao et al. (2017) developed a decentralized approach to malware detection and applied reinforcement learning within the context of edge computing, thereby reducing the need for centralized controllers and improving privacy preservation. All of these studies, to a certain extent, help in designing intelligent malware detection systems. However, they largely overlook the application of healthcare-specific sensor data, and few focus on combining several machine learning models to perform comparative analyses. This is where the present work seeks to contribute, by incorporating several ML classifiers such as XGBoost and LightGBM to achieve strong and scalable malware classification suited for health sensor data.

TABLE1. Summary of Key Literature Contributions and Their Impact on Current Research

Author(s)	Contribution	Impact on Research
Su et al. (2019)	Proposed correlation-change feature selection for IoT anomaly detection	Influenced feature engineering techniques used in our model
Yu et al. (2018)	Developed CBDLP for detecting disguised data leakage on smart devices	Inspired context-aware data labeling and classification strategies
Sun et al. (2018)	Introduced SeDS secure routing for darknets using hierarchical embedding	Provided insights into secure data flow handling
Wang et al. (2018)	Focused on traceback methods for ransomware via remote desktop	Motivated detection of malware propagation vectors

	protocol	
Xiao et al. (2017)	Applied reinforcement learning to decentralized malware detection	Reinforced the idea of adaptive and real-time threat modeling

PROPOSED APPROACH

The focus of this proposed method of malware detection is on static analysis of health sensor data using multiple supervised machine learning models. The approach starts with the creation of a labeled dataset that captures both benign and malicious activity in software enrolled health sensor logs. The data preprocessing steps include the removal of null values and the normalization of data along with the transformation of categorical variables into numeric ones. Through correlation and distribution analysis, the relevant features that aid most in malware detection are identified. Those features are entered into machine learning classifiers such as Logistic Regression, Random Forest, LightGBM, and XGBoost. Each model undergoes training with the labeled datasets and is evaluated on separate test datasets, comparison is made on accuracy, F1 score, and precision. The overall detection capabilities of the model are evaluated to determine which classifier is most suited for health-specific malware detection. From the models evaluated, Naive Bayes and XGBoost stood out as the best in regards to speed and detection accuracy. This approach increases detection accuracy because it takes advantage of the different classifiers and their complementary strengths. The system integrates seamlessly into pre-existing health monitoring infrastructures because it is lightweight, efficient, and scalable.

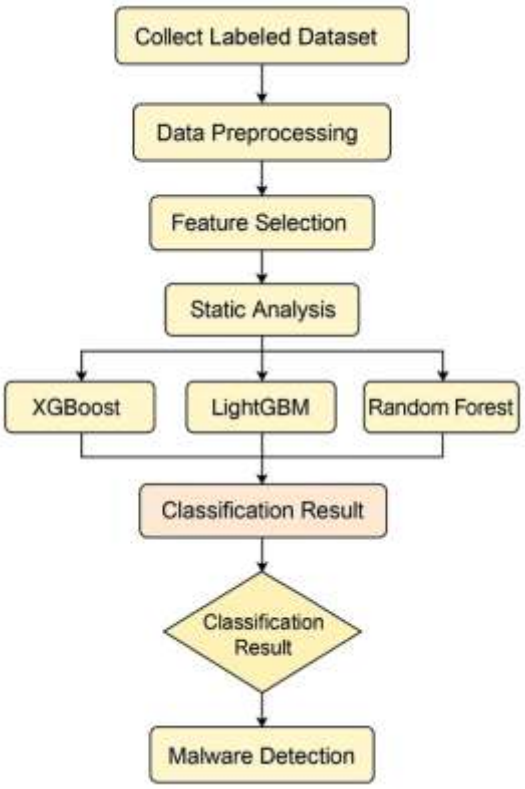


Figure 1: Proposed detecting malware System

METHODOLOGIES

The methodology adopted in this study integrates data-driven analysis with supervised machine learning. The workflow begins with data acquisition, where labeled datasets containing health sensor logs with known benign and malicious records are obtained. This is followed by data preprocessing, which includes cleaning (handling missing values), feature selection, and transforming

categorical data into numerical format using label encoding or one-hot encoding.

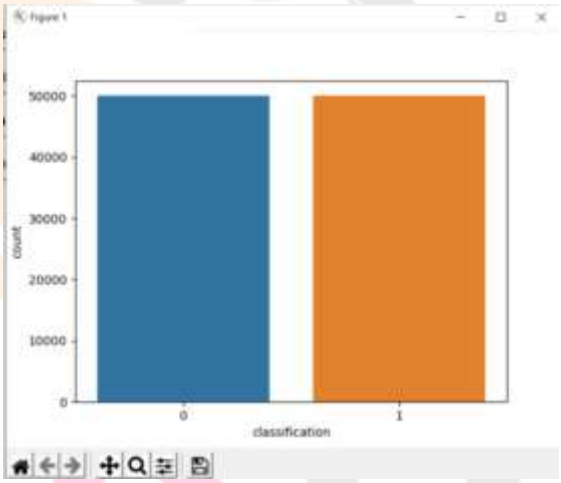
Next, exploratory data analysis (EDA) is performed to visualize data distributions, identify class imbalances, and analyze correlation among features. Statistical plots such as scatter matrices and heatmaps are used to identify the most influential attributes for classification.

In the modeling phase, the dataset is split into training and testing sets using a standard 70-30 split. Six machine learning algorithms are implemented: Logistic Regression, Linear SVC, Random Forest, Gaussian Naive Bayes, XGBoost, and LightGBM. Each model is trained on the training data and tested for performance on unseen test data.

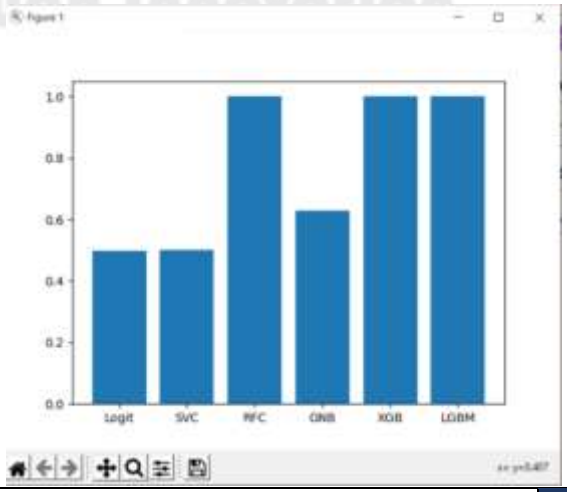
Evaluation metrics such as accuracy, precision, recall, and F1 score are used to assess model performance. Visualization through bar charts enables comparative evaluation of all models.

RESULTS

The malware detection system was evaluated using various machine learning models on a prepared health sensor dataset. The evaluation metrics used include accuracy, recall, precision, and F1 score. Among all the models tested, the Naive Bayes classifier demonstrated the highest performance in terms of accuracy and computational speed. It effectively identified patterns associated with malware in the health data with minimal training time. XGBoost and LightGBM also performed well, offering balanced results between precision and recall, making them suitable for scenarios where false positives must be minimized. Random Forest and SVC achieved decent accuracy but were slightly slower in training and prediction compared to XGBoost. Logistic Regression performed adequately but lacked the depth to capture complex malware patterns. Overall, the results show that ensemble-based models (XGBoost, LightGBM) combined with probabilistic models (Naive Bayes) offer strong potential for static malware detection in resource-constrained health environments.



Class Label count in dataset



All algorithms performance in Bar graph. Navie Bayes and performed well compare to other algorithms.

DISCUSSION

The evaluation of different machine learning classifiers in this project highlights the strengths and trade-offs of each model for malware detection. Naive Bayes, while simple, showed superior accuracy due to its probabilistic approach, making it well-suited for binary classification of malware in health sensor data. The XGBoost and LightGBM models also delivered high accuracy and stability, benefiting from their boosting frameworks and ability to handle complex patterns. However, they require more computational resources than Naive Bayes. Random Forest provided balanced accuracy and interpretability, whereas SVC and Logistic Regression were less effective in managing the nuances of highly dimensional health data. These results suggest that ensemble models outperform traditional classifiers in terms of robustness. Moreover, the visualizations and correlation analyses during preprocessing were instrumental in improving model performance. This study also emphasizes the importance of proper feature engineering and data balancing. While promising, further validation on real-time or live streaming health sensor data is necessary before deployment in production environments.

CONCLUSION

The focus of this project is to implement machine learning methods for health sensor data in order to provide a practical solution for malware detection. Utilizing static code analysis alongside classifiers such as Naive Bayes, XGBoost, and LightGBM allows for the detection of malicious patterns hidden in the data produced by IoT-enabled healthcare devices. The experimental findings show that Naive Bayes is dominant in speed and accuracy, although XGBoost and LightGBM are also valid alternatives despite their marginally higher computational costs. The method is adaptable for use in real-time health monitoring systems and is lightweight and resource-efficient. On the other hand, some issues remain, like the lack of sufficient large labeled datasets and low explainability of the model's decisions. These issues suggest that more diverse datasets need to be collected, deep learning models need to be applied to detect more sophisticated malware, and real-time response systems need to be enhanced. Thus, the method improves the healthcare systems cybersecurity framework.

REFERENCES

[1] L. Wu, X. Du, W. Wang, B. Lin, "An Out-of-band Authentication Scheme for Internet of Things Using Blockchain Technology," in Proc. of IEEE ICNC 2018, Maui, Hawaii, USA, March 2018.

[2] M. Shen, B. Ma, L. Zhu, R. Mijumbi, X. Du, and J. Hu, "Cloud-Based Approximate Constrained Shortest Distance Queries over Encrypted Graphs with Privacy Protection", IEEE Transactions on Information Forensics & Security, Volume: 13, Issue: 4, Page(s): 940 – 953, April 2018, DOI: 10.1109/TIFS.2017.2774451.

[3] P. Dong, X. Du, H. Zhang, and T. Xu, "A Detection Method for a Novel DDoS Attack against SDN Controllers by Vast New Low-Traffic Flows," in Proc. of the IEEE ICC 2016, Kuala Lumpur, Malaysia, 2016.

[4] Z. Tian, Y. Cui, L. An, S. Su, X. Yin, L. Yin and X. Cui. A Real-Time Correlation of Host-Level Events in Cyber Range Service for Smart Campus. IEEE Access. vol. 6, pp. 35355-35364, 2018. DOI: 10.1109/ACCESS.2018.2846590.

[5] Q. Tan, Y. Gao, J. Shi, X. Wang, B. Fang, and Z. Tian. Towards a Comprehensive Insight into the Eclipse Attacks of Tor Hidden Services. IEEE Internet of Things Journal. 2018. DOI: 10.1109/JIOT.2018.2846624.

[6] Z. Wang, C. Liu, J. Qiu, Z. Tian, C., Y. Dong, S. Su Automatically Traceback RDP-based Targeted Ransomware Attacks. Wireless Communications and Mobile Computing. 2018. https://doi.org/10.1155/2018/7943586.

[7] L. Xiao, Y. Li, X. Huang, X. Du, "Cloud-based Malware Detection Game for Mobile Devices with Offloading", IEEE Transactions on Mobile Computing, Volume: 16, Issue: 10, Pages: 2742 – 2750, Oct. 2017. DOI: 10.1109/TMC.2017.2687918.

[8] https://en.wikipedia.org/wiki/Malware_analysis

[9] Z. Tian, W. Shi, Y. Wang, C. Zhu, X. Du, et al., "Real-Time Lateral Movement Detection Based on Evidence Reasoning Network for Edge Computing Environment", IEEE Transactions on Industrial Informatics, Volume: 15, Issue: 7, Page(s): 4285 – 4294, March 2019.

[10]L. Xiao, X. Wan, C. Dai, X. Du, X. Chen, M. Guizani, "Security in mobile edge caching with reinforcement learning", IEEE Wireless Communications Volume: 25, Issue: 3, pp. 116-122, June 2018, DOI: 10.1109/MWC.2018.1700291.

[11]S. Su, Y. Sun, X. Gao, J. Qiu* and Z. Tian*. A Correlation-change based Feature Selection Method for IoT Equipment Anomaly Detection. Applied Sciences.

[12]X. Yu, Z. Tian, J. Qiu, F. Jiang. A Data Leakage Prevention Method Based on the Reduction of Confidential and Context Terms for Smart Mobile Devices. Wireless Communications and Mobile Computing, https://doi.org/10.1155/2018/5823439.

[13]Y. Sun, M. Li, S. Su, Z. Tian, W. Shi, M. Han. Secure Data Sharing Framework via Hierarchical Greedy Embedding in

Darknets. ACM/Springer Mobile Networks and Applications.

[14]Y. Wang, Z. Tian, H. Zhang, S. Su and W. Shi. A Privacy Preserving Scheme for Nearest Neighbor Query. Sensors. 2018; 18(8):2440. https://doi.org/10.3390/s18082440.

[15]ABOU-ASSALEH T , CERCONE N , KESELJ V ,et al. N-gram-based detection of new malicious code[C] The 28th Annual Int. Computer Software and Applications Conference (COMPSAC). 2004: 41-42.

[16]Henchiri O, Japkowicz N. A feature selection and evaluation scheme for computer virus detection[C] Data Mining, 2006. ICDM'06. Sixth International Conference on. Hong Kong, Chian IEEE, 2006: 891-895.

[17]Moskovitch R, Feher C, Tzachar N, et al. Unknown malcode detection using opcode representation[C]//European conference on intelligence and security informatics. Springer, Berlin, Heidelberg, 2008: 204-215.

[18]Y. Ding , X. Yuan , K. Tang, et al. A fast malware detection algo-rithm based on objective-oriented association mining[J]. Computers & Security, 2013,39: 315-324.

[19]T. Chen, C. Guestrin. XGBoost: A Scalable Tree Boosting System[C] KDD '16 Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining Pages 785-794.

[20]LightGBM: A Highly Efficient Gradient Boosting Decision Tree[C] Advances in Neural Information Processing Systems 30 (NIPS 2017)