Human Interaction and Emerging Technologies (IHIET 2026), Vol. 236, 2026, 473–482
https://doi.org/10.54941/ahfe1008122

Machine Learning-Based Prediction of Cybersecurity Vulnerability Severity for Enterprise Security Management

Iyad Suleiman and Adnan Agbaria

Department of Software Engineering, Kinneret Academic College, Israel

ABSTRACT

Cybersecurity vulnerabilities represent a growing challenge for modern information systems, particularly in large-scale cloud and enterprise environments where organizations must process a high volume of vulnerability disclosures under strict time constraints. Effective prioritization of security updates is essential for minimizing operational risk and improving resource allocation within Security Operations Centers (SOCs). However, traditional vulnerability severity assessment methods, such as the Common Vulnerability Scoring System (CVSS), rely heavily on manual analysis and expert judgment, limiting scalability in dynamic environments. This paper presents a machine learning–based framework for predicting the severity of cybersecurity vulnerabilities using structured data derived from Microsoft Security Bulletins. The dataset spans more than 15 years (2001–2017) and contains over 23,000 vulnerability records characterized by attributes such as impact type, affected product, affected component, and associated CVE identifiers. The prediction task is formulated as a multi-class classification problem with four severity levels: Critical, Important, Moderate, and Low. Several supervised learning models, including Logistic Regression, Decision Trees, Random Forest, and Gradient Boosting, are evaluated using macro-averaged precision, recall, and F1-score. Experimental results demonstrate that ensemble learning methods significantly outperform baseline classifiers. In particular, the Gradient Boosting model achieves the best performance, with a macro-F1 score of 0.982 and an overall accuracy of 99.0%. The findings demonstrate the effectiveness of machine learning techniques for automated vulnerability prioritization and highlight their practical applicability in enterprise cybersecurity management and SOC environments.

Keywords: Machine learning, Cyber security, Vulnerability, Enterprise security management

INTRODUCTION

The rapid growth of software complexity, cloud computing, and interconnected enterprise infrastructures has significantly increased the number and diversity of cybersecurity vulnerabilities. Organizations operating large-scale information systems must continuously monitor and manage a high volume of vulnerability disclosures, security alerts, and software updates. In such environments, timely and accurate prioritization of vulnerabilities is essential for minimizing operational risk, reducing potential attack surfaces, and

Received May 19, 2026; Revised June 27, 2026; Accepted July 22, 2026; Available online August 25, 2026

ensuring business continuity. However, determining which vulnerabilities require immediate remediation and which can be safely deferred remains a major challenge for Security Operations Centers (SOCs) and enterprise security teams.

To support vulnerability management processes, software vendors such as Microsoft publish detailed security bulletins containing structured information about vulnerability characteristics, affected products, exploit impact, and remediation recommendations. These bulletins provide valuable operational knowledge for assessing vulnerability severity and prioritizing patch deployment. Nevertheless, manual analysis of large volumes of security bulletins is time-consuming, labor-intensive, and potentially inconsistent, particularly in environments characterized by rapidly evolving threats and limited security resources.

Traditional vulnerability assessment approaches, including the Common Vulnerability Scoring System (CVSS), provide standardized mechanisms for estimating vulnerability severity. Although widely adopted, these methods rely heavily on expert judgment and manual metric assignment, limiting scalability and potentially introducing subjectivity into the evaluation process. Consequently, there is growing interest in applying machine learning techniques to automate vulnerability analysis and support data-driven cybersecurity decision-making.

In this paper, we propose a machine learning–based framework for predicting the severity of cybersecurity vulnerabilities using structured data derived from Microsoft Security Bulletins. The proposed approach formulates vulnerability severity estimation as a supervised multi-class classification problem and evaluates multiple machine learning models, including Logistic Regression, Decision Trees, Random Forest, and Gradient Boosting. The study utilizes a large-scale real-world dataset spanning more than 15 years of vulnerability disclosures, enabling a comprehensive evaluation under realistic operational conditions.

The experimental results demonstrate that ensemble learning methods significantly outperform traditional baseline models, achieving highly accurate severity predictions across all classes. From a practical perspective, the proposed framework can support automated vulnerability prioritization, improve SOC decision-making processes, and enhance resource allocation in enterprise cybersecurity environments.

RELATED WORK

Research on cybersecurity vulnerabilities has traditionally focused on identifying vulnerability types, attack vectors, and their impact on information systems. Early domain-specific studies, such as the work of Hugerat et al. (2013), analyzed vulnerabilities and attacks in e-learning environments, emphasizing the risks associated with delayed patching and inadequate vulnerability prioritization mechanisms. While these studies provided valuable insights into operational cybersecurity risks, they relied

primarily on manual analysis and did not provide automated mechanisms for predicting vulnerability severity. In contrast, the present study addresses this limitation by employing supervised machine learning techniques to enable automated severity prediction.

To standardize vulnerability severity assessment, the CVSS was introduced as a widely adopted framework for quantifying vulnerability impact and exploitability (Mell et al., 2007). CVSS provides a structured and consistent methodology for evaluating vulnerabilities and has become the de facto standard in enterprise cybersecurity environments. However, CVSS scoring depends heavily on manual metric assignment and expert interpretation, which may introduce subjectivity and limit scalability in environments characterized by large volumes of vulnerability disclosures. Unlike traditional CVSS-based approaches, the proposed framework leverages historical vulnerability data to automatically learn severity patterns, reducing reliance on manual scoring processes and enabling scalable vulnerability prioritization.

With the increasing availability of vulnerability datasets, machine learning techniques have been increasingly applied to automate vulnerability analysis and exploit prediction. (Bozorgi et al., 2010) demonstrated the feasibility of using supervised learning models to classify vulnerabilities and predict exploitability based on structured vulnerability attributes. Similarly, (Neuhaus et al., 2007) investigated the prediction of vulnerable software components using historical vulnerability data. These studies established the potential of data-driven approaches for cybersecurity risk assessment; however, they did not specifically address multi-class vulnerability severity prediction using large-scale, long-term datasets. In contrast, the present work formulates vulnerability severity prediction as a multi-class classification problem and evaluates multiple machine learning models using extensive real-world enterprise data.

More recent research has focused directly on machine learning–based vulnerability severity prediction. (Zhang et al., 2018) applied classification models to estimate CVSS-based severity levels and demonstrated that ensemble learning methods outperform traditional classifiers. In addition, (Allodi and Massacci, 2014) examined the relationship between vulnerability severity and real-world exploitation, highlighting the importance of accurate severity estimation for effective risk prioritization. Nevertheless, many existing studies rely primarily on publicly available datasets such as the National Vulnerability Database (NVD), often with limited temporal coverage or incomplete feature representations. In contrast, the present study utilizes a large-scale dataset derived from Microsoft Security Bulletins spanning more than 15 years, enabling a more comprehensive and operationally realistic evaluation of severity prediction models.

Despite the significant progress in machine learning–based vulnerability assessment, several challenges remain regarding dataset realism, scalability, and applicability to operational environments. Many existing approaches are evaluated using limited or partially curated datasets that may not fully

capture vendor-specific characteristics and long-term vulnerability trends. The proposed work addresses these limitations by leveraging a large vendor-specific dataset and focusing on performance metrics relevant to Security Operations Centers (SOCs), thereby bridging the gap between academic research and practical enterprise cybersecurity deployment.

DATA DESCRIPTION

The dataset used in this study was compiled from publicly available Microsoft Security Bulletin archives provided by the Microsoft Security Response Center. These bulletins contain structured information describing security vulnerabilities, affected products, and remediation details, offering a reliable and vendor-specific perspective on vulnerability assessment. Compared to commonly used public repositories such as the NVD, Microsoft Security Bulletins provide high-quality annotations that reflect real-world operational security evaluations in enterprise environments.

The dataset spans security updates published between 2001 and 2017 and includes approximately 23,500 vulnerability records. Each record corresponds to a documented security advisory and is represented by 14 structured features capturing both technical and contextual aspects of the vulnerability. The target variable is vulnerability severity, modeled as a multi-class categorical variable with four levels: Critical, Important, Moderate, and Low, following Microsoft's standardized severity rating system. The dataset includes attributes such as impact type, affected product, affected component, reboot requirement, number of associated Common Vulnerabilities and Exposures (CVE) identifiers, and publication date. These features collectively provide a comprehensive representation of vulnerability characteristics.

To ensure data quality and compatibility with machine learning algorithms, several preprocessing steps were applied. Duplicate records were removed to eliminate redundancy, and date fields were standardized to maintain temporal consistency across all entries. Missing values in categorical variables were handled using mode imputation, preserving the most frequent category and maintaining dataset integrity. Because the dataset contains multiple categorical variables, feature encoding techniques were applied to transform the data into a numerical representation suitable for supervised learning. Specifically, multi-class categorical features were encoded using one-hot encoding, while binary attributes, such as the reboot requirement indicator, were mapped to numerical values.

A notable characteristic of the dataset is the presence of class imbalance, with Critical and Important vulnerabilities occurring more frequently than Moderate and Low severity cases. This imbalance reflects real-world reporting trends but introduces challenges for model training, as classifiers may become biased toward majority classes. To address this issue, class weighting was incorporated during model training to penalize misclassification of minority classes. In addition, oversampling techniques, including the Synthetic Minority Over-sampling Technique, were evaluated to improve class balance and enhance predictive performance. The prediction task is formulated as a multi-class classification problem, where the objective is to estimate the severity level of a vulnerability based on its associated features. Several supervised learning models were implemented and evaluated, including Logistic Regression as a

baseline, Decision Tree, Random Forest, and Gradient Boosting within an XGBoost-style framework. This selection enables a comparative analysis across linear, tree-based, and ensemble learning approaches.

Given the imbalanced nature of the dataset, model performance was assessed using evaluation metrics that account for unequal class representation. These metrics include precision, recall, and macro-average F1-score. Macro-F1 was selected as the primary evaluation metric because it assigns equal importance to all classes, providing a balanced assessment of performance across severity levels.

For experimental evaluation, the dataset was divided into training and testing subsets using a stratified 80/20 split to preserve class distribution. To enhance robustness and reduce variance, five-fold cross-validation was applied during model training. Hyperparameter optimization for ensemble-based models was conducted using grid search over key parameters, including tree depth, number of estimators, and learning rate, with the objective of maximizing predictive performance while mitigating overfitting.

OUR METHODOLOGY

The objective of this study is to develop a predictive model for estimating the severity level of cybersecurity vulnerabilities based on structured attributes extracted from security bulletins. The problem is formulated as a supervised multi-class classification task. Let x∈R^(n×d) denote the feature matrix representing n vulnerability instances with d attributes, and let y∈{1,2,3,4} denote the corresponding severity labels, representing the four classes: Critical, Important, Moderate, and Low. The goal is to learn a mapping function f:x→y that accurately predicts the severity class of unseen vulnerabilities.

The proposed methodology follows a standard machine learning pipeline consisting of feature representation, model training, and performance evaluation. Given the structured nature of the dataset, multiple supervised learning models are employed to capture different types of relationships between vulnerability attributes and severity levels. Logistic Regression is used as a baseline model due to its simplicity and interpretability, providing a linear decision boundary for classification. Decision Trees are employed to model non-linear relationships through recursive partitioning of the feature space. To further improve predictive performance, ensemble learning techniques are incorporated. Random Forest aggregates multiple decision trees trained on different subsets of the data to reduce variance and improve generalization. Gradient Boosting constructs an ensemble sequentially by iteratively minimizing prediction errors, enabling the model to capture complex patterns and subtle feature interactions.

The choice of these models is motivated by their effectiveness in handling structured data and their ability to model heterogeneous feature types. Ensemble methods are well suited for cybersecurity applications, where vulnerability severity is influenced by multiple interacting factors that exhibit non-linear dependencies.

Hyperparameter optimization is performed for the ensemble models using grid search over key parameters such as tree depth, number of estimators, and learning rate. This process aims to identify configurations that maximize

predictive performance while reducing the risk of overfitting. To account for class imbalance, class-weighted learning is incorporated into the training process, ensuring that minority classes are adequately represented during optimization.

Overall, the proposed methodology enables a systematic comparison between linear, tree-based, and ensemble learning approaches, providing insight into their relative effectiveness for vulnerability severity prediction.

RESULTS

This section presents the experimental evaluation of the proposed machine learning models for vulnerability severity prediction using the Microsoft Security Bulletin dataset. The models are evaluated on a held-out test set using macro-averaged precision, recall, and F1-score, which provide a balanced assessment of performance across all severity classes.

After preprocessing and deduplication, the dataset is divided into training and testing subsets using a stratified 80/20 split, resulting in 15,252 training samples and 3,813 test samples (19,065 records in total). As shown in Table 1, the class distribution is highly imbalanced, with the Important and Critical classes dominating the dataset, while Low severity vulnerabilities represent less than 1% of the total instances. This imbalance motivates the use of macro-average metrics and class-weight training strategies.

Table 1: Class distribution across training and test splits.

Severity Class	Train Count	Train %	Test Count	Test %
Critical	5,596	36.7%	1,448	38.0%
Important	8,226	53.9%	1,993	52.3%
Low	127	0.8%	43	1.1%
Moderate	1,303	8.5%	329	8.6%
Total	15,252	100%	3,813	100%

Table 2 summarizes the overall performance of the evaluated models. The results reveal a clear performance hierarchy. Logistic Regression, serving as a linear baseline, achieves a macro-F1 score of 0.428, indicating limited capability in capturing complex relationships within the data. The Decision Tree model significantly improves performance, achieving a macro-F1 score of 0.917 by modeling non-linear decision boundaries.

Table 2: Macro-averaged performance metrics on the test set.

Model	Macro Precision	Macro Recall	Macro F1-Score	Accuracy
Logistic Regression	0.443	0.612	0.428	55.1%
Decision Tree	0.886	0.954	0.917	93.3%
Random Forest	0.979	0.967	0.973	98.3%
Gradient Boosting	0.990	0.975	0.982	99.0%

Ensemble-based models further enhance predictive performance. Random Forest achieves a macro-F1 score of 0.973, demonstrating strong generalization and robustness across all severity classes. The Gradient Boosting model achieves the best overall performance, with a macro-F1 score of 0.982 and an accuracy of 99.0%. These results confirm that ensemble learning methods are highly effective for vulnerability severity prediction.

A detailed per-class analysis, presented in Table 3, provides additional insights into model behavior. Critical and Important classes are consistently predicted with high accuracy across ensemble models, with F1-scores exceeding 0.98. Importantly, the Low severity class, despite being severely underrepresented, is also effectively handled by Random Forest and Gradient Boosting, achieving F1-scores of 0.963 and 0.977, respectively. This demonstrates the effectiveness of class-weighted training in mitigating the impact of class imbalance. In contrast, Logistic Regression performs poorly on minority classes, with an F1-score of 0.113 for Low severity and 0.375 for Moderate severity.

Table 3: Per-class classification results. LR = Logistic Regression, DT = Decision Tree, RF = Random Forest, GB = Gradient Boosting.

Model	Class	Precision	Recall	F1-Score	Support
LR	Critical	0.598	0.443	0.509	1,448
LR	Important	0.832	0.627	0.715	1,993
LR	Low	0.062	0.814	0.113	43
LR	Moderate	0.282	0.562	0.375	329
DT	Critical	0.921	0.931	0.926	1,448
DT	Important	0.967	0.926	0.946	1,993
DT	Low	0.843	1.000	0.915	43
DT	Moderate	0.813	0.961	0.881	329
RF	Critical	0.983	0.983	0.983	1,448
RF	Important	0.991	0.993	0.992	1,993
RF	Low	0.976	0.953	0.963	43
RF	Moderate	0.967	0.939	0.953	329
HGB	Critical	0.989	0.989	0.989	1,448
HGB	Important	0.991	0.991	0.991	1,993
HGB	Low	1.000	0.953	0.977	43
HGB	Moderate	0.982	0.967	0.975	329

Confusion matrix analysis (Figure 1) further highlights differences between models. Logistic Regression exhibits substantial misclassification between adjacent severity levels, particularly between Critical, Important, and Moderate classes. Decision Trees reduce these errors but still show confusion near class boundaries. Random Forest significantly improves classification consistency, reducing cross-class misclassification. Gradient Boosting achieves the most accurate classification, with predictions concentrated along the diagonal and minimal confusion between severity levels.

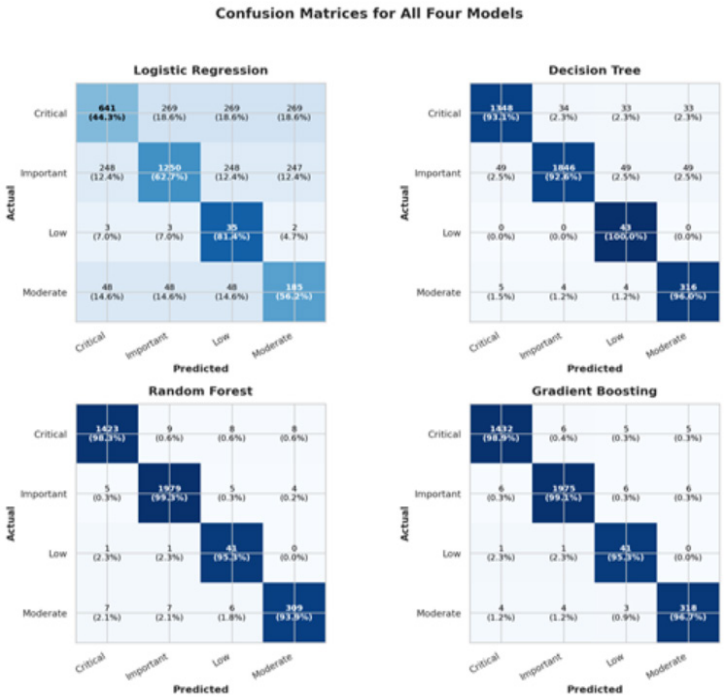

Figure 1: Confusion matrices for all four models. Cell color encodes row-normalised rate; raw counts are annotated. Gradient Boosting shows the most concentrated diagonal.

Feature importance analysis derived from the Random Forest model reveals that impact type, affected product, and bulletin-related attributes are the most influential predictors of vulnerability severity. These features collectively account for a substantial portion of the model's predictive power, indicating that severity is primarily driven by the technical impact and context of vulnerability. In contrast, features such as the number of associated CVEs and reboot requirements contribute less to model performance.

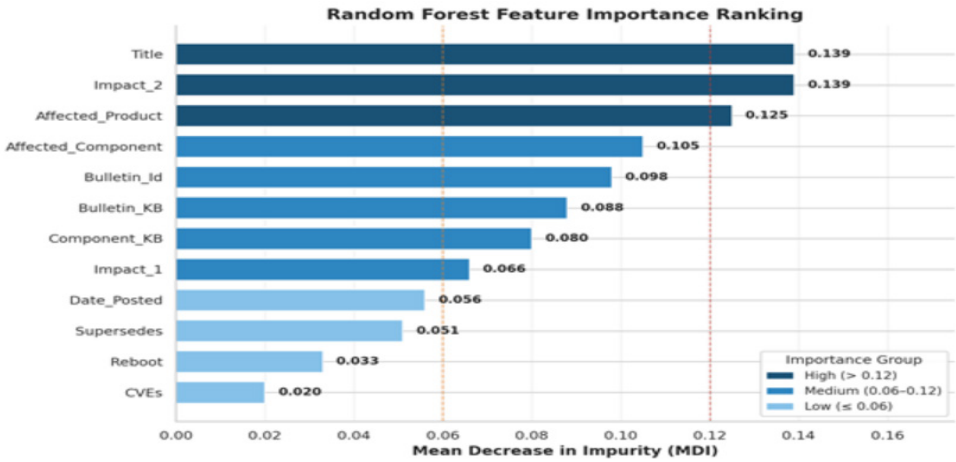

Figure 2: Feature importance ranking from the Random Forest classifier. Bulletin title, secondary impact type, and affected product are the strongest predictors of severity.

Overall, the results demonstrate that ensemble learning models, particularly Gradient Boosting, provide highly accurate and robust predictions for vulnerability severity. The strong performance across both majority and minority classes highlights the effectiveness of the proposed approach and supports its applicability in real-world Security Operations Center (SOC) environments for automated vulnerability prioritization.

CONCLUSION AND FUTURE WORK

This paper presented a machine learning–based approach for predicting the severity of cybersecurity vulnerabilities using structured data derived from Microsoft Security Bulletins. The problem was formulated as a multi-class classification task, and multiple supervised learning models were evaluated, including Logistic Regression, Decision Trees, Random Forest, and Gradient Boosting.

The experimental results demonstrate that ensemble learning methods significantly outperform traditional linear models. In particular, the Gradient Boosting model achieved the best performance, reaching a macro-F1 score of 0.982 and providing highly accurate predictions across all severity classes. Importantly, strong performance was maintained even for minority classes, indicating the effectiveness of class-weighted learning in addressing dataset imbalance. These findings confirm that vulnerability severity can be reliably predicted using structured features extracted from security bulletins.

The results also highlight the practical relevance of the proposed approach. By enabling automated severity prediction, the model can support Security Operations Centers (SOCs) in prioritizing vulnerability remediation efforts, reducing analyst workload, and improving response efficiency in large-scale enterprise environments.

Despite these promising results, several limitations should be acknowledged. The study relies on historical data covering the period up to 2017 and does not incorporate real-time threat intelligence or behavioral indicators. Additionally, while ensemble models provide strong predictive performance, they may reduce interpretability, which is an important consideration for operational deployment.

Future work will focus on extending the proposed framework in several directions. First, incorporating more recent vulnerability data and streaming security feeds could improve model adaptability to emerging threats. Second, integrating temporal and contextual features, such as CVSS temporal metrics or exploit availability, may enhance prediction accuracy. Third, the application of explainable artificial intelligence (XAI) techniques could improve transparency and support decision-making by security analysts. Finally, exploring hybrid models that combine structured features with textual analysis of vulnerability descriptions represents a promising direction for further research.

ACKNOWLEDGMENT

The authors would like to thank Rozan Halabi and Fahid Jamoly for their contributions to the analysis.

REFERENCES

Allodi, L., & Massacci, F. (2014). Comparing vulnerability severity and exploits using case-control studies. ACM Transactions on Information and System Security (TISSEC), 17(1), 1–20.

Bozorgi, M., Saul, L. K., Savage, S., & Voelker, G. M. (2010). Beyond heuristics: Learning to classify vulnerabilities and predict exploits. Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 105–114.

Hugerat, M., Saker, S., Odeh, S., and Agbaria, A. (2013). Vulnerabilities and Attacks on Information Systems in E-learning Environments in Higher Education. US-China Education Review A, ISSN 2161-623X. Vol. 3, No. 8, 615–622.

Mell, P., Scarfone, K., & Romanosky, S. (2007). A complete guide to the Common Vulnerability Scoring System (CVSS). FIRST—Forum of Incident Response and Security Teams.

Neuhaus, S., Zimmermann, T., Holler, C., & Zeller, A. (2007). Predicting vulnerable software components. Proceedings of the 14th ACM Conference on Computer and Communications Security (CCS), 529–540.

Zhang, Z., Wang, Y., Liu, Y., & Li, J. (2018). Vulnerability severity prediction based on machine learning. Security and Communication Networks, 2018, Article 1960981.