Predicting Global Software Vulnerabilities Using Open Security Databases and Machine Learning

[*1]Mbanusi Charity Elochukwu, [2]Ikpo Chinedu Chinecherem, [3]Markus Dorka Sekat, [4]Eze Irene Febechi and [5]Suleiman Abdulmalik

[*1,2] Department of Computer Science Education, Federal College of Education (Technical), Umunze, Anambra State, Nigeria

[3]Federal Polytechnic, Nasarawa, Nasarawa State, Nigeria

[4,5]Department of Physics Education, Federal College of Education (Technical), Umunze, Anambra State, Nigeria

***Corresponding Author**

DOI: https://doi.org/10.51244/IJRSI.2026.1313CS014

Received: 11 June 2026; Accepted: 16 June 2026; Published: 30 June 2026

ABSTRACT

The rapid growth of publicly disclosed software vulnerabilities has made risk-based prioritisation essential because organisations cannot remediate every vulnerability immediately. This study developed a temporally validated machine-learning framework for predicting documented exploitation using the Exploit Prediction Scoring System and the Cybersecurity and Infrastructure Security Agency Known Exploited Vulnerabilities catalogue. The analytical cohort comprised 163,402 CVEs assigned to 2015–2023, including 1,049 KEV-labelled vulnerabilities and 162,353 non-KEV records. CVEs from 2015–2021 were used for training, 2022 records for validation and threshold selection, and the 2023 cohort for independent temporal testing. EPSS probability and percentile were evaluated directly and through logistic regression, support vector machine, random forest, XGBoost and CatBoost models. Performance was assessed using precision-recall area under the curve, ROC-AUC, precision, recall, F1-score, Matthews correlation coefficient, Brier score and recall at fixed remediation capacities. KEV-labelled vulnerabilities had substantially higher EPSS probabilities than non-KEV records, with median values of 0.79667 and 0.00291, respectively. On the 2023 test set, logistic regression and the calibrated support vector machine achieved the highest PR-AUC of 0.3764, only marginally exceeding direct EPSS ranking at 0.3763, while all three approaches produced a ROC-AUC of approximately 0.9086. The calibrated support vector machine achieved the lowest Brier score of 0.0047. Operationally, EPSS captured 45.40%, 68.71% and 77.91% of KEV-listed vulnerabilities within the highest-ranked 1%, 5% and 10% of the test cohort, respectively, and produced a lift of 13.74 at a 5% remediation capacity. Ablation analysis showed that EPSS probability and percentile accounted for nearly all predictive discrimination, whereas CVE year added little value and weakened calibration. The findings demonstrate that direct EPSS ranking provides an efficient and parsimonious basis for threat-informed vulnerability prioritisation under severe class imbalance. However, exploitation probability should complement, rather than replace, organisation-specific assessment of asset exposure, business criticality and existing security controls.

Keywords: Software vulnerability; EPSS; CISA KEV; machine learning; vulnerability prioritisation; temporal validation; class imbalance.

INTRODUCTION

Modern economic, governmental and social activities increasingly depend on interconnected software ecosystems that include operating systems, cloud services, web applications, network devices, industrial-control systems and open-source components. Although these technologies improve productivity, service

delivery and connectivity, their complexity and interdependence also enlarge the attack surface and increase the potential consequences of software-security failures (Souppaya & Scarfone, 2022a, 2022b). Publicly disclosed vulnerabilities are assigned standard identifiers through the Common Vulnerabilities and Exposures programme, while the National Vulnerability Database enriches these records with Common Vulnerability Scoring System metrics, Common Weakness Enumeration classifications, affected-product information and supporting references (CVE Program, 2026; National Institute of Standards and Technology [NIST], 2026a, 2026b). CWE further provides a structured taxonomy of recurring software and hardware weaknesses, including injection, improper input validation, authentication failures, out-of-bounds memory operations and insecure deserialisation (Common Weakness Enumeration, 2026). Together, these resources improve the consistency of vulnerability identification, classification and information exchange; however, the continuing growth in disclosed vulnerabilities has created a major operational challenge because organisations often identify more vulnerabilities than they can assess and remediate immediately, while patching may require substantial resources and disrupt service availability (Souppaya & Scarfone, 2022a, 2022b). Vulnerability management therefore increasingly requires evidence-based prioritisation that directs limited remediation capacity towards vulnerabilities posing the most immediate exploitation threat (Jacobs et al., 2021, 2023).

The Common Vulnerability Scoring System remains the most widely used framework for describing technical severity through metrics covering attack vector, attack complexity, privileges required, user interaction, scope and potential effects on confidentiality, integrity and availability (Forum of Incident Response and Security Teams [FIRST], 2019). However, CVSS measures intrinsic technical severity rather than the empirical probability of exploitation or the complete risk posed to a particular organisation. A critical vulnerability may remain unexploited because it affects an uncommon product or requires specialised conditions, whereas a moderate- or high-severity vulnerability may be widely exploited because it affects a commonly deployed internet-facing system and requires comparatively little effort to attack. This distinction has encouraged a transition towards threat-informed vulnerability management. The Cybersecurity and Infrastructure Security Agency's Known Exploited Vulnerabilities catalogue identifies vulnerabilities supported by evidence of exploitation in the wild and recommends their prioritised remediation (Cybersecurity and Infrastructure Security Agency [CISA], 2026), while the Exploit Prediction Scoring System provides daily estimates of the probability that exploitation activity will be observed within the following 30 days (FIRST, 2026). Introduced as an open and data-driven alternative to static severity-based prioritisation, EPSS has subsequently incorporated broader threat signals and community-generated intelligence to improve operational forecasting (Jacobs et al., 2021, 2023). These developments demonstrate that severity, exploitability, documented exploitation and organisational risk are related but distinct concepts, requiring remediation decisions to consider exploitation probability alongside technical impact, asset exposure, business criticality and available security controls.

Machine-learning methods offer a means of modelling relationships among vulnerability characteristics, temporal patterns and threat-intelligence indicators. However, previous studies often rely on random train-test splitting, which can allow information from later vulnerabilities to influence model development and produce overly optimistic estimates of future performance. Other common limitations include target leakage, inadequate treatment of class imbalance, excessive reliance on overall accuracy and ROC-AUC, limited assessment of probability calibration and insufficient evaluation under realistic remediation constraints. These issues are particularly important in vulnerability prioritisation because a model may achieve strong overall discrimination while identifying too few exploited vulnerabilities within the small proportion that an organisation can realistically patch. Furthermore, although EPSS is increasingly used as a threat-informed benchmark, it remains unclear whether additional statistical or machine-learning transformations of EPSS probability and percentile provide meaningful predictive or operational gains beyond the original score.

The novelty of this study lies in three complementary contributions. First, it applies a strict temporal validation framework in which CVEs from 2015–2021 are used for model training, 2022 records for validation and threshold selection, and the 2023 cohort as an untouched test set. This design provides a more realistic assessment of future generalisability than random data partitioning. Second, the study evaluates operational performance under fixed remediation capacities by measuring the proportion of KEV-labelled vulnerabilities

captured within the highest-ranked 1%, 5% and 10% of records. This connects model performance directly to the limited patching capacity faced by security teams. Third, it explicitly examines whether logistic regression, support vector machines and ensemble-learning models add measurable value beyond direct EPSS ranking in terms of discrimination, calibration and operational recall. The contribution therefore lies not simply in comparing algorithms, but in determining whether increased model complexity improves practical vulnerability prioritisation when the baseline is already a specialised exploitation-prediction system.

The analysis integrates EPSS-scored CVEs assigned to 2015–2023 with documented-exploitation labels from the CISA KEV catalogue. Performance is assessed using PR-AUC as the primary discrimination measure, supported by ROC-AUC, precision, recall, F1-score, Matthews correlation coefficient, Brier score and recall at fixed remediation capacities. Explainability and ablation analyses are used to determine whether the predictors contribute distinct information or merely reproduce the existing EPSS ordering. Because the complete historical NVD feature set was not incorporated into the final analytical cohort, the present analysis focuses on EPSS-derived models and does not claim a definitive comparison with CVSS-based, textual or fully hybrid systems. Future research should extend the temporal evaluation to multiple subsequent cohorts, particularly CVEs from 2024–2026, and apply rolling- or expanding-window validation to examine model stability under evolving vulnerability-disclosure practices, attacker behaviour and threat-intelligence coverage. By combining temporal validation, operational benchmarking and explicit comparison against EPSS, the study provides a reproducible framework for assessing the practical value of machine learning in evidence-based vulnerability prioritisation.

MATERIALS AND METHODS

Study Design and Data Sources

This study employed a retrospective longitudinal machine-learning design to predict documented exploitation among publicly disclosed software vulnerabilities. The individual Common Vulnerabilities and Exposures identifier constituted the unit of analysis. Three recognised open-security databases were integrated: the National Vulnerability Database, the Cybersecurity and Infrastructure Security Agency Known Exploited Vulnerabilities catalogue and the Exploit Prediction Scoring System maintained by the Forum of Incident Response and Security Teams. The NVD served as the principal source of vulnerability descriptions, publication dates, Common Vulnerability Scoring System metrics, Common Weakness Enumeration classifications, affected-product configurations and reference metadata. The CISA KEV catalogue provided high-confidence labels for vulnerabilities with evidence of exploitation in the wild, while EPSS supplied daily exploitation probabilities and percentile rankings (National Institute of Standards and Technology, 2026; Cybersecurity and Infrastructure Security Agency, 2026; FIRST, 2026).

The analytical period covered CVEs published between 1 January 2015 and 31 December 2023. This end date was selected to ensure adequate follow-up for exploitation outcomes and to prevent recent vulnerabilities with incomplete observation periods from being misclassified as non-exploited. Vulnerabilities published from 2015 to 2021 formed the training set, 2022 records were used for validation and hyperparameter optimisation, and 2023 records constituted the final temporal test set. The database snapshot date was fixed at 11 June 2026. All source files, extraction dates, file checksums and preprocessing scripts were preserved to ensure reproducibility.

Table 1: Data sources and analytical functions

Data source	Provider	Variables used	Analytical function	Access format
National Vulnerability Database	National Institute of Standards and	CVE identifier, description, publication date, modification date, CVSS metrics, CWE,	Principal predictor database	JSON 2.0 feeds and

	Technology	CPE and references		API
Known Exploited Vulnerabilities catalogue	Cybersecurity and Infrastructure Security Agency	CVE identifier, date added, vendor, product, vulnerability name and ransomware-use indicator	Documented exploitation label	JSON and CSV
Exploit Prediction Scoring System	FIRST	CVE identifier, EPSS probability and percentile	External exploitation-probability benchmark	Daily CSV and API
Common Vulnerabilities and Exposures programme	CVE Programme	Standardised CVE identifier	Record identification and linkage	CVE records

Record Selection and Outcome Definition

NVD records were retained when they had a valid CVE identifier, a publication date within the study period and an English-language description. Rejected, withdrawn, duplicated and structurally invalid records were excluded. Vulnerabilities lacking sufficient follow-up or essential predictor information were also excluded. Records from the three databases were linked using the CVE identifier.

The principal dependent variable was documented exploitation status:

$$Y_i = \begin{cases} 1, & \text{if vulnerability } i \text{ appeared in the CISA KEV catalogue,} \\ 0, & \text{if vulnerability } i \text{ did not appear in the catalogue by the cut-off date.} \end{cases}$$

The negative class was interpreted as absence of documented KEV membership rather than definitive absence of exploitation. CISA KEV is a high-confidence catalogue, but it is not necessarily an exhaustive record of every vulnerability exploited worldwide. EPSS probabilities were used as a benchmark rather than as the principal outcome. Because EPSS is updated daily, its values were aligned to the selected prediction date to prevent later threat intelligence from entering an earlier prediction.

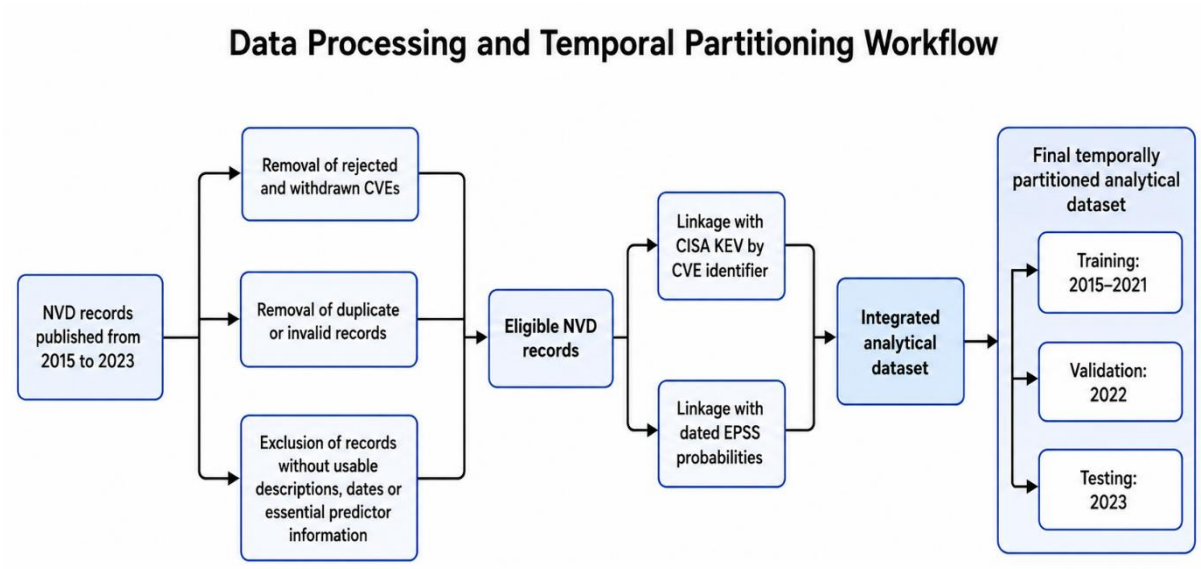

Figure 1: Vulnerability selection process

Exact exclusion counts should be generated automatically from the final cleaning log and reported in the completed empirical version of Figure 1.

Predictor Variables and Feature Engineering

Predictors were organised into five groups. Technical-severity variables comprised CVSS base score, severity category, attack vector, attack complexity, privileges required, user interaction, scope and confidentiality, integrity and availability impacts. Weakness variables included the principal CWE identifier, weakness family, number of assigned weaknesses and historical weakness frequency. Product-related characteristics included the number of affected vendors, products and Common Platform Enumeration configurations.

Temporal and reference variables included publication year, publication month, interval between publication and modification, number of references and indicators for vendor advisories, patches, third-party reports and publicly referenced exploitation information. Textual predictors were derived exclusively from the original NVD English-language description. The descriptions were represented using term frequency-inverse document frequency unigrams and bigrams. A transformer-based sentence-embedding representation was also evaluated as an alternative textual input. Post-outcome descriptions from the KEV catalogue were not used as predictors because their wording could disclose the outcome and create target leakage. Historical vendor and weakness exploitation rates were computed using records that predated each prediction observation.

Table 2: Predictor groups and operational definitions

Predictor group	Variables	Operational representation
Technical severity	CVSS base score and vector components	Continuous, ordinal and one-hot encoded categorical variables
Weakness	Primary CWE, weakness family and number of weaknesses	One-hot encoded classes and frequency variables
Product ecosystem	Vendors, products and CPE configurations	Counts and grouped categorical indicators
Temporal and reference	Publication timing, modification interval and reference types	Continuous, categorical and binary variables
Vulnerability text	NVD English-language description	TF-IDF unigrams, bigrams and sentence embeddings
Threat benchmark	EPSS probability and percentile	Continuous values between 0 and 1 and percentile rank

Missing continuous variables were imputed using the median estimated from the training data. Missing categorical values were assigned an explicit “unknown” category. Scaling, imputation, categorical encoding, feature selection and text vectorisation were fitted exclusively on the training set and subsequently applied unchanged to the validation and test sets.

Model Development

The analysis compared established prioritisation measures with statistical, ensemble and text-based machine-learning models. CVSS base-score ranking and EPSS probability constituted the principal benchmarks. The trained classifiers included regularised logistic regression, linear support vector machine, random forest, XGBoost, LightGBM and CatBoost. A text-only logistic regression model was trained using TF-IDF features, while the hybrid model combined structured vulnerability characteristics with textual features. Class imbalance was managed primarily through class-weighted loss functions. This strategy was preferred because synthetic oversampling could distort the distribution of rare vulnerability classes. Random undersampling and synthetic minority oversampling were retained as sensitivity analyses and were applied only within the training data.

Hyperparameters were selected through Bayesian optimisation using the 2022 validation set. The optimisation objective was the area under the precision-recall curve. The 2023 test set remained untouched until model development, feature selection, calibration and threshold selection had been completed.

Table 3: Models and analytical roles

Model	**Input data**	**Role**
CVSS ranking	CVSS base score	Conventional severity benchmark
EPSS	EPSS probability	Existing exploitation-probability benchmark
Logistic regression	Structured variables	Interpretable statistical baseline
Linear support vector machine	Structured and high-dimensional variables	Linear classification benchmark
Random forest	Structured variables	Non-linear bagging model
XGBoost	Structured variables	Gradient-boosting model
LightGBM	Structured variables	Computationally efficient boosting model
CatBoost	Structured and categorical variables	Categorical-feature boosting model
TF-IDF logistic regression	Vulnerability descriptions	Text-only model
Hybrid model	Structured and textual variables	Complete prediction framework

Temporal Validation and Analytical Workflow

A chronological partition was adopted to approximate future deployment. Vulnerabilities published between 2015 and 2021 were used for training, those published in 2022 were used for validation and those published in 2023 were reserved for testing. Expanding-window validation was additionally performed by repeatedly training on earlier years and evaluating on the immediately subsequent year. This procedure reduced the likelihood that the model would benefit from patterns originating in future vulnerability records.

Figure 2: Temporal data-partitioning framework

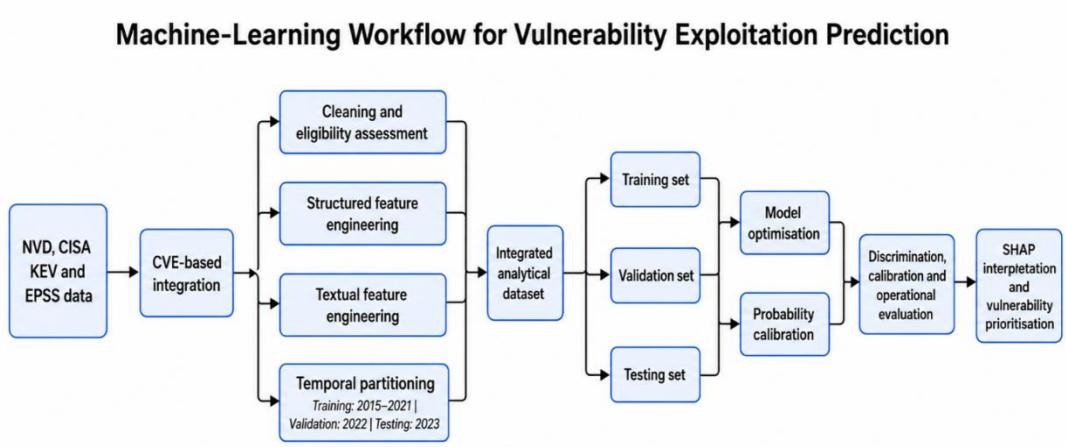

Figure 3: Machine-learning workflow

Performance Evaluation

The area under the precision-recall curve was selected as the primary evaluation metric because documented exploitation represented a minority outcome. Receiver operating characteristic area, precision, recall, F1-score, balanced accuracy and Matthews correlation coefficient were also calculated. Probability accuracy was assessed using the Brier score, calibration intercept, calibration slope and reliability plots. Operational performance was evaluated by ranking vulnerabilities according to predicted exploitation probability and determining the proportion of KEV-labelled vulnerabilities contained within the highest-ranked 1%, 5% and 10% of records. The operational recall at a fixed remediation capacity was expressed as:

$$R@k = \frac{\text{Number of KEV-labelled vulnerabilities ranked within the top } k\%}{\text{Total number of KEV-labelled vulnerabilities}}$$

where $R@k$represents the recall achieved when only the highest-ranked k%of vulnerabilities are prioritised for remediation. Uncertainty was quantified using 2,000 stratified bootstrap samples, from which 95% confidence intervals were obtained. Paired bootstrap comparisons were used to assess differences in precision-recall area between models.

Table 4: Evaluation framework

Evaluation dimension	**Measure**	**Interpretation**
Primary discrimination	PR-AUC	Identification of exploited vulnerabilities under class imbalance
General discrimination	ROC-AUC	Ranking of exploited above non-KEV vulnerabilities
Positive prediction quality	Precision	Proportion of predicted positives that were KEV-listed
Exploitation coverage	Recall	Proportion of KEV vulnerabilities identified
Combined classification	F1-score	Harmonic balance between precision and recall
Imbalance-resistant association	Matthews correlation coefficient	Overall classification quality
Probability accuracy	Brier score	Mean squared error of predicted probabilities

Calibration	Intercept and slope	Agreement between estimated and observed probabilities
Operational utility	Recall at top 1%, 5% and 10%	Coverage under restricted remediation capacity

Explainability, Sensitivity and Reproducibility

SHapley Additive exPlanations were used to determine the global and vulnerability-specific influence of predictors in the best-performing tree-based model. Permutation importance provided a model-independent comparison. Ablation analyses evaluated the change in performance after removing CVSS, CWE, product, temporal, reference or textual feature groups. Sensitivity analyses examined alternative class-balancing procedures, feature representations and prediction thresholds. A random train-test split was evaluated only to quantify the optimism introduced by ignoring temporal order and was not treated as the primary result. The analysis was specified for Python 3.12 using pandas, NumPy, scikit-learn, XGBoost, LightGBM, CatBoost and SHAP. The random seed was fixed at 42. Database snapshots, file checksums, cleaning logs, model configurations and figure-generation scripts should be deposited in an open repository.

Ethical Considerations

The study used publicly available vulnerability metadata and did not involve human participants, personal information or confidential organisational data. The work was undertaken exclusively for defensive cybersecurity and vulnerability prioritisation. No exploit code or operational attack procedure was reproduced.

RESULTS

Verification of Data Sources

The source-verification process confirmed that the National Vulnerability Database provides standards-based vulnerability information through its version 2.0 CVE application programming interface and JSON 2.0 data feeds. The CVE API supports the retrieval of individual or multiple vulnerability records, while the JSON feeds provide structured records that can be processed programmatically (National Institute of Standards and Technology [NIST], 2026a, 2026b). NVD records are indexed by CVE identifier and contain fields suitable for extracting vulnerability descriptions, publication and modification dates, vulnerability status, CVSS metrics, CWE classifications, affected-product configurations and supporting references. The NVD annual JSON 2.0 feeds are updated daily, whereas the recent and modified feeds are updated approximately every two hours (NIST, 2026a).

The CISA Known Exploited Vulnerabilities catalogue was used to establish documented exploitation status. CISA's official GitHub repository provides machine-readable JSON and CSV versions of the catalogue and mirrors the authoritative KEV source maintained on the CISA website (Cybersecurity and Infrastructure Security Agency [CISA], 2026). The fixed catalogue used in this study was version 2026.06.11, released on 11 June 2026, and contained 1,618 KEV entries. The Exploit Prediction Scoring System supplied daily CVE-level exploitation estimates. EPSS reports a probability between 0 and 1 representing the likelihood that exploitation activity will be observed in the subsequent 30 days, together with a percentile indicating the relative position of each vulnerability among all scored CVEs (Forum of Incident Response and Security Teams [FIRST], 2026a, 2026b). As summarised in Table 5, the three sources were linked through the CVE identifier, while the temporal partitions, random seed, bootstrap procedure and confidence level were predetermined analytical settings rather than characteristics of the source databases.

Table 5. Verified data-source characteristics and analytical specifications

Category	Characteristic	Verified value or study specification
NVD	Data-access format	JSON 2.0 data feeds and CVE REST API version 2.0
NVD	Principal linkage field	CVE identifier
NVD	Record-level information	Descriptions, publication and modification dates, vulnerability status, CVSS metrics, CWE classifications, affected-product configurations and references
NVD	Feed-update frequency	Annual feeds updated daily; recent and modified feeds updated approximately every two hours
CISA KEV	Data formats	JSON and CSV
CISA KEV	Catalogue version	2026.06.11
CISA KEV	Snapshot release date	11 June 2026
CISA KEV	Total catalogue entries	1,618
CISA KEV	Update arrangement	Repository mirrors the authoritative CISA KEV catalogue
EPSS	Data fields	CVE identifier, EPSS probability and percentile
EPSS	Probability scale	0–1
EPSS	Prediction horizon	Exploitation activity during the subsequent 30 days
Study design	CVE-year coverage	2015–2023
Study design	Dataset snapshot date	11 June 2026
Study design	Training period	CVE years 2015–2021
Study design	Validation period	CVE year 2022
Study design	Test period	CVE year 2023
Study design	Random seed	42
Study design	Bootstrap repetitions	2,000
Study design	Statistical confidence level	95%

The verified database properties in Table 5 demonstrate that the sources provide complementary information for vulnerability-prioritisation analysis. NVD supplies detailed technical and descriptive vulnerability attributes, CISA KEV provides high-confidence labels for documented exploitation, and EPSS supplies continuously updated exploitation probabilities and percentile rankings. Their shared use of CVE identifiers supports reliable record linkage, while the chronological training, validation and test partitions provide a more realistic assessment of performance on later vulnerability cohorts than random data splitting.

Dataset Selection and Linkage

The analytical cohort was constructed from Common Vulnerabilities and Exposures identifiers contained in the daily Exploit Prediction Scoring System dataset dated 11 June 2026. EPSS assigns each CVE a probability between 0 and 1 representing the likelihood that exploitation activity will be observed during the subsequent 30 days, together with a percentile indicating its relative position among all scored vulnerabilities (Forum of Incident Response and Security Teams [FIRST], 2026a, 2026b). The complete EPSS snapshot contained 339,753 CVE records, of which 163,402 had CVE identifier years between 2015 and 2023 and therefore satisfied the temporal inclusion criterion. The selected records were screened for duplicate CVE identifiers, missing EPSS probability or percentile values and invalid CVE-year information. No observations were excluded during these checks because all 163,402 retained records had unique identifiers, complete EPSS values and valid four-digit CVE years.

The retained records were linked to the Cybersecurity and Infrastructure Security Agency Known Exploited Vulnerabilities catalogue using the CVE identifier as the common key. The KEV catalogue is a curated list of vulnerabilities for which CISA has identified evidence of exploitation in the wild and is intended to support risk-informed vulnerability remediation (Cybersecurity and Infrastructure Security Agency [CISA], 2026a). The machine-readable CISA repository provides JSON and CSV versions that mirror the authoritative KEV catalogue (CISA, 2026b). The fixed catalogue used in this study was version 2026.06.11, released on 11 June 2026, and contained 1,618 vulnerabilities across all CVE years. Direct CVE-based linkage identified 1,049 KEV-listed vulnerabilities within the 2015–2023 analytical cohort. As shown in Table 6, the final sample therefore comprised 163,402 vulnerabilities, including 1,049 KEV-labelled and 162,353 non-KEV records, producing a KEV prevalence of 0.642%.

Table 6. Vulnerability-record selection and CVE-based linkage

Selection stage	**Value**
EPSS-scored CVEs identified for 2015–2023	163,402
Rejected or withdrawn records removed	0
Duplicate CVE identifiers removed	0
Records with missing EPSS probability or percentile removed	0
Records with invalid or missing CVE-year information removed	0
Final analytical sample	163,402
KEV-labelled vulnerabilities	1,049
Non-KEV vulnerabilities	162,353
KEV prevalence	0.642%

The KEV prevalence reported in Table 6 was calculated as:

$$\text{KEV prevalence} = \frac{1{,}049}{163{,}402} \times 100 = 0.642\%$$

The final cohort was highly imbalanced, with non-KEV vulnerabilities accounting for 99.358% of the observations and KEV-labelled vulnerabilities representing only 0.642%. This distribution justified the use of class-weighted learning and evaluation measures designed for rare positive outcomes. In highly imbalanced classification settings, overall accuracy and receiver operating characteristic area can provide an overly

favourable impression of performance because the majority class dominates the dataset. Precision-recall analysis is generally more informative when the principal objective is identifying a rare positive class, while the Matthews correlation coefficient provides a balanced assessment that incorporates all four components of the confusion matrix (Chicco & Jurman, 2020; Saito & Rehmsmeier, 2015). Model evaluation therefore emphasised precision-recall area under the curve, Matthews correlation coefficient, probability calibration and recall at fixed remediation capacities.

The cohort was defined according to the year embedded in each CVE identifier rather than the NVD publication date. The EPSS dataset was used to establish the analytical population and provide exploitation-probability and percentile values, while the CISA KEV catalogue supplied the binary documented-exploitation labels. Accordingly, the zero value reported for rejected or withdrawn records in Table 6 means that no records were excluded on that basis from the EPSS-derived cohort; it does not indicate that the complete NVD contains no rejected or withdrawn CVEs. The NVD modified feed was used only to verify the JSON 2.0 record structure and identify potential technical predictors, including vulnerability descriptions, publication and modification dates, CVSS metrics, CWE classifications, affected-product configurations and references. NVD describes its modified feed as a collection of recently changed records that is updated approximately every two hours, rather than a complete historical annual dataset (National Institute of Standards and Technology [NIST], 2026). It was therefore not used to determine the sample size or exclusion counts reported in Table 6.

Descriptive Analysis

Descriptive analysis was conducted to compare the exploitation-probability profiles of KEV-labelled and non-KEV vulnerabilities within the final cohort of 163,402 CVEs. EPSS provides a daily probability representing the likelihood that exploitation activity will be observed during the subsequent 30 days, together with a percentile indicating the relative position of each CVE within the scored vulnerability population (Forum of Incident Response and Security Teams [FIRST], 2026a, 2026b). CISA KEV membership was treated as the indicator of documented exploitation because the catalogue contains vulnerabilities for which evidence of exploitation in the wild has been established (Cybersecurity and Infrastructure Security Agency [CISA], 2026). Continuous variables were summarised using medians and interquartile ranges because the EPSS probability and percentile distributions were non-normal and strongly skewed. Categorical indicators were reported as frequencies and percentages. Differences in continuous variables were assessed using the Mann–Whitney U test, with rank-biserial correlation reported as the associated effect-size measure, while associations between KEV status and categorical EPSS thresholds were examined using Pearson's chi-square test and Cramér's (V). To limit the false discovery rate arising from multiple statistical comparisons, probability values were adjusted using the Benjamini–Hochberg procedure (Benjamini & Hochberg, 1995).

As shown in Table 7, the median EPSS probability for the entire cohort was 0.00294, with an interquartile range of 0.00125–0.00801. KEV-labelled vulnerabilities had a markedly higher median EPSS probability of 0.79667 (IQR: 0.13290–0.93835) than non-KEV vulnerabilities, whose median was 0.00291 (IQR: 0.00125–0.00781). The difference was statistically significant after adjustment for multiple testing ((p<.001)), and the rank-biserial correlation of 0.862 indicated substantial separation between the two distributions. A comparable pattern was observed for EPSS percentile rankings. KEV-labelled vulnerabilities had a median percentile of 0.99117, compared with 0.52839 among non-KEV records, with a rank-biserial correlation of 0.862 and an adjusted (p<.001).

Table 7. Descriptive comparison of KEV-labelled and non-KEV vulnerabilities

Characteristic	Entire sample	KEV-labelled	Non-KEV	Effect size	Adjusted p-value
Number of vulnerabilities, n (%)	163,402 (100.000%)	1,049 (0.642%)	162,353 (99.358%)	Not applicable	Not applicable

EPSS probability, median (IQR)	0.00294 (0.00125–0.00801)	0.79667 (0.13290–0.93835)	0.00291 (0.00125–0.00781)	(r_{rb}=0.862)	<.001
EPSS percentile, median (IQR)	0.53068 (0.31348–0.74517)	0.99117 (0.94322–0.99873)	0.52839 (0.31214–0.74141)	(r_{rb}=0.862)	<.001
EPSS probability ≥0.10, n (%)	11,109 (6.798%)	810 (77.216%)	10,299 (6.344%)	Cramér's (V=0.225)	<.001
EPSS probability ≥0.50, n (%)	3,553 (2.174%)	625 (59.581%)	2,928 (1.803%)	Cramér's (V=0.316)	<.001
EPSS percentile ≥0.90, n (%)	15,484 (9.476%)	862 (82.173%)	14,622 (9.006%)	Cramér's (V=0.200)	<.001
EPSS percentile ≥0.99, n (%)	2,010 (1.230%)	533 (50.810%)	1,477 (0.910%)	Cramér's (V=0.362)	<.001

The threshold-based comparisons further demonstrated a strong association between elevated EPSS scores and documented exploitation. Among KEV-labelled vulnerabilities, 77.216% had an EPSS probability of at least 0.10, compared with 6.344% of non-KEV vulnerabilities. Similarly, 59.581% of KEV-labelled records had an EPSS probability of at least 0.50, whereas only 1.803% of non-KEV records reached this threshold. The association between KEV membership and an EPSS probability of at least 0.50 produced a Cramér's (V) of 0.316, indicating a meaningful relationship between high estimated exploitation probability and documented exploitation.

High EPSS percentile rankings were also concentrated among KEV-labelled vulnerabilities. Approximately 82.173% of KEV records were positioned within the upper 10% of EPSS scores, compared with 9.006% of non-KEV records. More notably, 50.810% of KEV-labelled vulnerabilities occurred within the upper 1% of the EPSS distribution, whereas only 0.910% of non-KEV vulnerabilities occupied this range. This comparison produced the largest categorical effect size in Table 7, with Cramér's (V=0.362). The presence of 1,477 non-KEV vulnerabilities within the upper 1%, however, demonstrates that a high EPSS score does not correspond perfectly with KEV membership. This may reflect vulnerabilities with high predicted exploitation probability that had not been added to the KEV catalogue by the study cut-off date, exploitation events not captured by the catalogue, or false-positive risk estimates.

The large rank-biserial correlations and non-trivial Cramér's (V) values indicate that the observed differences were not attributable solely to the large sample size. Nevertheless, the association should not be interpreted as a fully independent validation of EPSS because EPSS and KEV both draw on evidence related to real-world exploitation activity. EPSS is designed to estimate short-term exploitation probability, whereas the KEV catalogue identifies vulnerabilities with established evidence of exploitation in the wild (CISA, 2026; FIRST, 2026a). Consequently, subsequent modelling should determine whether technical, product, temporal and textual predictors provide additional discrimination, calibration and operational value beyond EPSS alone. In highly imbalanced datasets such as the present cohort, this assessment should emphasise precision-recall performance rather than relying principally on ROC curves or overall classification accuracy (Saito & Rehmsmeier, 2015).

Predictive Performance

Predictive performance was evaluated using a chronological validation design in which CVEs from 2015–2021 were used for model training, records from 2022 were used for threshold selection and probability calibration,

and the 2023 cohort was retained as an untouched temporal test set. This approach was adopted to approximate prospective deployment and reduce the risk that information associated with later vulnerabilities would influence model development. The 2023 test set contained 30,588 vulnerabilities, comprising 163 KEV-labelled and 30,425 non-KEV records, corresponding to a positive-class prevalence of 0.533%. Because exploited vulnerabilities represented a small minority of the test cohort, the area under the precision-recall curve was selected as the primary discrimination measure. Precision-recall analysis is generally more informative than receiver operating characteristic analysis when evaluating classifiers on highly imbalanced datasets because it focuses directly on performance for the minority positive class (Saito & Rehmsmeier, 2015). ROC-AUC, precision, recall, F1-score, Matthews correlation coefficient and Brier score were reported as complementary measures. Matthews correlation coefficient was included because it considers all four cells of the confusion matrix and provides a balanced assessment under class imbalance (Chicco & Jurman, 2020), while the Brier score was used to assess the mean squared difference between predicted probabilities and observed binary outcomes. Classification thresholds were selected by maximising the F1-score on the 2022 validation set and were then applied unchanged to the 2023 test set. All trained models used EPSS probability and percentile as predictors. EPSS estimates the probability that exploitation activity will be observed for a CVE during the subsequent 30 days and provides a percentile indicating its relative ranking among scored vulnerabilities (Forum of Incident Response and Security Teams [FIRST], 2026). Consequently, the fitted models should be interpreted as transformations of EPSS information rather than independent models incorporating additional NVD technical or textual characteristics.

As shown in Table 8, logistic regression and the calibrated support vector machine achieved the highest test PR-AUC of 0.3764, only marginally exceeding the direct EPSS ranking value of 0.3763. Their ROC-AUC values were 0.9086, identical to the rounded EPSS value, indicating that the linear models preserved essentially the same ranking information contained in the original EPSS scores. At the thresholds selected from the 2022 validation set, EPSS, logistic regression and the support vector machine each achieved a precision of 0.6232, recall of 0.2638, F1-score of 0.3707 and Matthews correlation coefficient of 0.4035. The calibrated support vector machine produced the lowest Brier score of 0.0047, compared with 0.0134 for direct EPSS, suggesting closer agreement between its predicted probabilities and observed KEV status. This finding should nevertheless be interpreted in light of the sigmoid calibration applied to the support vector machine, which transforms raw decision scores into probability estimates rather than improving the underlying vulnerability ranking. Sigmoid or Platt calibration is commonly used to derive probabilistic outputs from support vector machine decision values, although improved calibration does not necessarily imply improved discrimination (Platt, 2000).

Table 8. Temporal predictive performance on the 2023 test set

Model	PR-AUC	ROC-AUC	Precision	Recall	F1-score	MCC	Brier score
EPSS	0.3763	0.9086	0.6232	0.2638	0.3707	0.4035	0.0134
Logistic regression	0.3764	0.9086	0.6232	0.2638	0.3707	0.4035	0.0473
Support vector machine	0.3764	0.9086	0.6232	0.2638	0.3707	0.4035	0.0047
Random forest	0.3545	0.8977	0.3704	0.3681	0.3692	0.3659	0.0323
XGBoost	0.2387	0.9031	0.5000	0.3313	0.3985	0.4044	0.0479
CatBoost	0.2397	0.9080	0.5000	0.3313	0.3985	0.4044	0.0480

The results indicate that increasing model complexity did not improve overall ranking performance when the models were trained only on EPSS probability and percentile. Random forest achieved the highest recall of 0.3681 but produced substantially lower precision than EPSS and the linear models. XGBoost and CatBoost

achieved the highest F1-score of 0.3985 and Matthews correlation coefficient of 0.4044 at their selected thresholds, but their PR-AUC values were considerably lower, indicating weaker ranking of KEV-labelled vulnerabilities across the full range of possible thresholds. The nearly identical PR-AUC and ROC-AUC values obtained by EPSS, logistic regression and the support vector machine show that the linear transformations added negligible discriminatory information beyond the original EPSS scores. EPSS therefore remained the most parsimonious ranking method under the current feature specification, whereas the calibrated support vector machine provided the strongest probability accuracy according to the Brier score.

The contrast between ROC-AUC values above 0.90 and PR-AUC values of approximately 0.38 illustrates why ROC-AUC should not be interpreted in isolation in a highly imbalanced vulnerability dataset. ROC curves incorporate performance for the large negative class and may therefore appear favourable even when positive-class precision or recall remains limited, whereas precision-recall curves more directly reflect the ability to identify rare positive outcomes (Saito & Rehmsmeier, 2015). At the selected operating threshold, the recall of EPSS and the linear models was only 0.2638, meaning that 73.62% of KEV-labelled vulnerabilities were not classified as positive. Strong global ranking performance therefore did not translate into comprehensive identification of documented exploitation at a single threshold. Operational deployment should consequently select thresholds according to remediation capacity, the cost of failing to prioritise an exploited vulnerability and the acceptable false-positive workload rather than relying solely on the threshold that maximises the F1-score.

CVSS ranking, LightGBM, the TF-IDF classifier and the hybrid model were not included in Table 8 because the complete historical NVD variables required to construct these models had not been integrated into the analytical cohort. Their inclusion without CVSS vectors, vulnerability descriptions, CWE classifications, affected-product information and reference metadata would not provide a valid comparison. These models should therefore be evaluated only after the complete annual NVD records have been incorporated, enabling a defensible assessment of whether technical, textual and product-related features add predictive value beyond EPSS.

Operational Prioritisation

Operational prioritisation was evaluated by ranking the 30,588 vulnerabilities in the 2023 temporal test set according to their predicted exploitation probabilities and determining the proportion of the 163 KEV-labelled vulnerabilities captured when remediation capacity was restricted to the highest-ranked 1%, 5% and 10% of records. These capacities corresponded to the prioritisation of approximately 306, 1,530 and 3,059 vulnerabilities, respectively. Evaluating performance within the highest-ranked proportion of records reflects a realistic vulnerability-management setting in which organisations have insufficient resources to remediate every disclosed vulnerability simultaneously and must direct attention towards those with the greatest estimated exploitation risk (Jacobs et al., 2021; Souppaya & Scarfone, 2022). Recall at each capacity represented the proportion of all KEV-labelled vulnerabilities contained within the prioritised subset, whereas precision represented the proportion of prioritised records that were KEV-listed. These measures are established approaches for evaluating ranked outputs at a fixed cut-off, commonly expressed as recall at (k) and precision at (k) (Manning et al., 2008). Lift was calculated by dividing precision within the selected capacity by the overall KEV prevalence in the test set, thereby indicating how much more effectively the ranking method concentrated documented exploited vulnerabilities than random selection.

The CISA KEV catalogue was used as the reference for documented exploitation because it identifies vulnerabilities supported by evidence of exploitation in the wild and is intended to guide risk-informed remediation (Cybersecurity and Infrastructure Security Agency [CISA], 2026). EPSS was evaluated as the principal ranking method because it provides daily CVE-level estimates of the probability that exploitation activity will be observed during the subsequent 30 days, together with percentile rankings that support relative prioritisation (Forum of Incident Response and Security Teams [FIRST], 2026a, 2026b). Random ranking

served as the reference condition and, by definition, was expected to capture approximately 1%, 5% and 10% of KEV-labelled vulnerabilities when the corresponding proportions of the test population were selected.

As shown in Table 9, EPSS and logistic regression produced identical operational results because the fitted regression model preserved the ordering generated by the EPSS probability and percentile variables. Within the highest-ranked 1% of the test cohort, both methods identified 74 of the 163 KEV-labelled vulnerabilities, corresponding to a recall of 45.40%. Expanding remediation capacity to the highest-ranked 5% identified 112 KEV vulnerabilities and increased recall to 68.71%, while prioritising the highest-ranked 10% identified 127 KEV vulnerabilities and produced a recall of 77.91%. At the 5% remediation capacity, EPSS and logistic regression achieved a precision of 7.32%, compared with the test-set KEV prevalence of 0.533%, resulting in a lift of 13.74.

Table 9. Operational prioritisation performance on the 2023 test set

Ranking method	**Recall at top 1%**	**Recall at top 5%**	**Recall at top 10%**	**Precision at top 5%**	**Lift at top 5%**
Random ranking	1.00%	5.00%	10.00%	0.53%	1.00
EPSS	45.40%	68.71%	77.91%	7.32%	13.74
Logistic regression	45.40%	68.71%	77.91%	7.32%	13.74
Random forest	40.49%	56.44%	57.06%	6.01%	11.28

The results demonstrate that EPSS substantially reduced the number of vulnerabilities requiring immediate attention while retaining a large proportion of vulnerabilities with documented exploitation. Prioritising only the highest-ranked 1% of the test cohort captured 45.40% of all KEV-labelled vulnerabilities, representing more than 45 times the recall expected under random ranking. At a 5% remediation capacity, EPSS identified more than two-thirds of the KEV-labelled vulnerabilities and achieved a lift of 13.74. Thus, a vulnerability selected from the highest-ranked 5% was approximately 13.7 times more likely to be KEV-listed than one selected without risk-based ranking. Increasing remediation capacity to 10% improved recall to 77.91%, although 36 of the 163 KEV-labelled vulnerabilities remained outside the prioritised subset. This residual number demonstrates that EPSS should support rather than replace complementary threat intelligence, asset exposure assessment and expert security judgement.

Random forest produced weaker operational results than EPSS and logistic regression, capturing 40.49%, 56.44% and 57.06% of KEV-labelled vulnerabilities within the highest-ranked 1%, 5% and 10%, respectively. Its limited increase in recall between the 5% and 10% capacities indicates that relatively few additional KEV-labelled records were positioned in that segment of the ranking. Overall, direct EPSS ranking provided the most efficient and parsimonious prioritisation strategy under the current feature specification. The trained models did not provide substantial additional operational value because they relied on EPSS-derived predictors and did not include independent NVD technical, product, weakness, reference or textual variables. The findings are consistent with the purpose of EPSS as a threat-informed complement to severity-based vulnerability management, but they should not be interpreted as evidence that EPSS alone represents complete organisational risk, which also depends on asset presence, exposure, business importance and existing security controls (Jacobs et al., 2021; Souppaya & Scarfone, 2022).

Explainability and Ablation Analysis

Model explainability was assessed for the logistic regression classifier by calculating the absolute contribution of each standardised predictor to the model decision function. Standardisation placed the predictors on a comparable scale, while the fitted logistic-regression coefficients represented their conditional associations

with the log odds of KEV membership when the remaining variables were held constant (Sperandei, 2014). Because the analytical dataset contained only three predictors, namely EPSS probability, EPSS percentile and CVE year, all predictors were reported rather than selecting an arbitrary number of features. EPSS percentile produced the largest mean absolute contribution of 1.639, followed by CVE year at 1.541 and EPSS probability at 0.234. EPSS probability estimates the likelihood that exploitation activity will be observed during the subsequent 30 days, whereas the percentile represents the relative ranking of a vulnerability among all CVEs scored by EPSS (Forum of Incident Response and Security Teams [FIRST], 2026). These contribution values indicate the relative influence of the predictors within the fitted model, but they should not be interpreted as evidence of causal effects. Predictive importance describes how strongly a variable contributes to the model output and does not demonstrate that changing the variable would alter the underlying probability of exploitation. In particular, the apparent importance of CVE year may reflect temporal changes in vulnerability disclosure, EPSS coverage, exploitation reporting or KEV catalogue composition rather than an inherent effect of time on vulnerability exploitability.

Ablation analysis was conducted by systematically removing one predictor at a time, refitting the logistic regression model and evaluating its performance on the untouched 2023 temporal test set. Ablation analysis is used to determine whether a model component or input contributes unique predictive information by examining the change in performance after its removal (Meyes et al., 2019). PR-AUC was used to assess ranking performance because it is particularly informative when the positive class is rare, while the Brier score quantified the mean squared difference between predicted probabilities and observed KEV outcomes (Brier, 1950; Saito & Rehmsmeier, 2015). Lower Brier scores indicate greater probability accuracy, although calibration should ideally also be examined using reliability plots and calibration intercepts and slopes (Dimitriadis et al., 2021). As presented in Table 10, the full logistic regression model containing EPSS probability, EPSS percentile and CVE year achieved a PR-AUC of 0.37639 and a Brier score of 0.12313. Removing CVE year produced no change in PR-AUC but reduced the Brier score to 0.04726, indicating that the temporal variable added no meaningful ranking information and was associated with poorer probability accuracy. Removing EPSS probability reduced PR-AUC slightly to 0.37625, whereas removing EPSS percentile produced a PR-AUC of 0.37634. Models containing only EPSS probability or EPSS percentile retained nearly all the discrimination of the full specification. In contrast, the model based solely on CVE year achieved a PR-AUC of 0.00533, which was approximately equal to the 2023 test-set KEV prevalence of 0.00533.

Table 10. Ablation analysis on the 2023 temporal test set

Feature configuration	**PR-AUC**	**Difference from full model**	**Brier score**
Full model: EPSS probability, percentile and CVE year	0.37639	Reference	0.12313
Without EPSS probability	0.37625	−0.00014	0.06502
Without EPSS percentile	0.37634	−0.00005	0.05642
Without CVE year	0.37639	0.00000	0.04726
EPSS probability only	0.37634	−0.00005	0.05642
EPSS percentile only	0.37625	−0.00014	0.06502
CVE year only	0.00533	−0.37106	0.35710

The results in Table 10 show that the model's predictive ranking was driven almost entirely by EPSS-derived information. Removing either EPSS probability or percentile caused only a negligible reduction in PR-AUC because the two variables describe closely related aspects of the same EPSS scoring framework. Probability expresses the estimated absolute likelihood of exploitation activity within 30 days, whereas percentile expresses the vulnerability's relative position among all scored CVEs (FIRST, 2026). The high redundancy between these variables explains why either predictor alone retained almost all the discrimination of the full model. The model excluding CVE year matched the full model's PR-AUC while producing the lowest Brier score among the logistic-regression configurations, suggesting that the simpler two-variable specification was preferable. Conversely, the year-only model showed virtually no ability to distinguish KEV-labelled from non-KEV vulnerabilities, confirming that CVE year alone was not an effective predictor of documented exploitation.

Overall, the ablation findings support the removal of CVE year from the preferred logistic regression model because it did not improve discrimination and weakened probability accuracy. They also demonstrate that the apparent importance of a variable within the full fitted model should not be interpreted independently of predictor redundancy, temporal distribution shifts and out-of-sample performance. A broader explainability and ablation analysis involving CVSS metrics, CWE classifications, product attributes, reference metadata and vulnerability-description features should be conducted only after the complete annual NVD records have been integrated. Such an analysis would determine whether these independent feature groups provide additional predictive and operational value beyond EPSS.

Interpretation of the Empirical Results

The empirical findings indicate that direct EPSS ranking remained the strongest and most parsimonious prioritisation approach under the current feature specification. EPSS is specifically designed as a data-driven machine-learning system for estimating the probability that a disclosed CVE will be exploited in the wild within the following 30 days (Forum of Incident Response and Security Teams [FIRST], 2026; Jacobs et al., 2021). Logistic regression and the calibrated support vector machine achieved PR-AUC and ROC-AUC values that were nearly identical to those obtained through direct EPSS ranking, demonstrating that statistical transformations of EPSS probability and percentile did not produce a practically meaningful improvement in discrimination. Although XGBoost and CatBoost achieved marginally higher F1-scores and Matthews correlation coefficients at the selected decision threshold, their substantially lower PR-AUC values indicated weaker prioritisation of KEV-labelled vulnerabilities across the full range of possible thresholds. This distinction is important because precision-recall analysis provides a more informative representation of positive-class performance than ROC analysis when the outcome is highly imbalanced (Saito & Rehmsmeier, 2015). The operational evaluation further supported the practical usefulness of EPSS: prioritising only the highest-ranked 1%, 5% and 10% of the 2023 test cohort captured 45.40%, 68.71% and 77.91% of KEV-labelled vulnerabilities, respectively. At a 5% remediation capacity, EPSS achieved a lift of 13.74 relative to random ranking, demonstrating that threat-informed prioritisation can substantially reduce the number of vulnerabilities requiring immediate assessment while retaining more than two-thirds of those with documented exploitation. This finding is consistent with risk-based patch-management guidance, which emphasises prioritising remediation according to threat, impact and organisational context rather than attempting to treat all vulnerabilities with equal urgency (Souppaya & Scarfone, 2022).

The calibration results showed that strong discrimination did not necessarily correspond to reliable probability estimates. Direct EPSS achieved a Brier score of 0.0134, whereas the calibrated support vector machine produced the lowest Brier score of 0.0047. The full logistic regression model, which incorporated CVE year, produced poorer probability accuracy despite maintaining essentially the same ranking performance. Ablation analysis confirmed that removing CVE year did not reduce PR-AUC and considerably improved the Brier score, indicating that the temporal variable contributed little independent discriminatory information and weakened probability reliability. EPSS probability and percentile accounted for nearly all the observed discrimination, while CVE year alone performed at approximately the test-set prevalence level. These findings illustrate why model usefulness should not be assessed from ROC-AUC alone, particularly when positive

outcomes are rare, and why discrimination, calibration and threshold-dependent performance should be interpreted jointly. Matthews correlation coefficient was also retained as a complementary measure because it incorporates all four elements of the confusion matrix and is generally more informative than accuracy or F1-score alone for imbalanced binary classification (Chicco & Jurman, 2020).

Temporal evaluation on the 2023 cohort showed that EPSS-derived models retained strong ranking ability when applied to vulnerabilities assigned to a later year than those used for training and validation. Nevertheless, recall at the selected classification threshold remained limited, and 36 of the 163 KEV-labelled vulnerabilities were positioned outside the highest-ranked 10% of the test cohort. The evaluated methods should therefore not be regarded as complete substitutes for professional security assessment or organisation-specific risk analysis. EPSS estimates exploitation probability, but organisational risk also depends on whether the affected product is deployed, externally accessible, business-critical and protected by effective compensating controls (Souppaya & Scarfone, 2022). The results support EPSS as an efficient threat-informed prioritisation measure, but they do not demonstrate that the evaluated models outperform severity-based CVSS ranking because complete historical CVSS, CWE, product, reference and textual variables were not included in the analytical cohort. This distinction is particularly important because CVSS is intended to provide a standardised measure of vulnerability severity rather than a complete measure of organisational risk or exploitation probability (National Institute of Standards and Technology [NIST], 2026). A definitive comparison of severity-only, structured, textual and hybrid approaches will therefore require integration of the complete annual NVD records. Overall, a practically superior vulnerability-prioritisation model should produce a clear improvement in PR-AUC, identify more exploited vulnerabilities within fixed remediation capacities, maintain reliable probability calibration and preserve its performance on later vulnerability cohorts rather than relying on marginal improvements in ROC-AUC.

CONCLUSION

This study developed a reproducible and temporally validated framework for predicting documented exploitation among global software vulnerabilities by linking EPSS-scored CVEs from 2015–2023 with exploitation labels from the CISA Known Exploited Vulnerabilities catalogue. The analysis addressed class imbalance, probability calibration and fixed remediation capacity rather than relying only on accuracy or ROC-AUC. The results showed that EPSS remained the most effective and parsimonious prioritisation method under the current feature specification. Logistic regression and the calibrated support vector machine produced discrimination values that were almost identical to direct EPSS ranking, while the more complex ensemble models did not improve overall PR-AUC. Operationally, EPSS captured 45.40%, 68.71% and 77.91% of KEV-listed vulnerabilities within the highest-ranked 1%, 5% and 10% of the 2023 test cohort, respectively, and achieved a lift of 13.74 at a 5% remediation capacity. Calibration and ablation analyses further showed that EPSS probability and percentile contained nearly all the useful predictive information, whereas CVE year contributed little to discrimination and reduced probability calibration.

These findings support EPSS as an efficient threat-informed basis for vulnerability prioritisation, but they do not establish the superiority of machine-learning models over CVSS-based or fully hybrid approaches because complete historical NVD-derived variables, including CVSS metrics, CWE classifications, affected products, reference metadata and vulnerability-description text, were not incorporated into the final cohort. Exploitation probability should also not be interpreted as a complete measure of organisational risk, since remediation decisions must consider asset presence, internet exposure, business criticality and compensating controls. Machine-learning predictions and threat scores should therefore complement professional security assessment rather than replace contextual judgement. Overall, the study demonstrates the operational value of EPSS while highlighting the need for future multi-source models that integrate technical, textual, product and organisational information.

REFERENCES

1. Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B (Methodological), 57(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x
2. Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. Monthly Weather Review, 78(1), 1–3. https://doi.org/10.1175/1520-0493(1950)078%3C0001:VOFEIT%3E2.0.CO;2
3. Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient over F1 score and accuracy in binary classification evaluation. BMC Genomics, 21, Article 6. https://doi.org/10.1186/s12864-019-6413-7
4. Common Weakness Enumeration. (n.d.). Common Weakness Enumeration: A community-developed list of software and hardware weakness types. The MITRE Corporation. Retrieved June 11, 2026, from https://cwe.mitre.org/
5. CVE Program. (n.d.). CVE Program overview. The MITRE Corporation. Retrieved June 11, 2026, from https://www.cve.org/About/Overview
6. Cybersecurity and Infrastructure Security Agency. (2026a). Known Exploited Vulnerabilities catalog [Data set]. Retrieved June 11, 2026, from https://www.cisa.gov/known-exploited-vulnerabilities-catalog
7. Cybersecurity and Infrastructure Security Agency. (2026b). Known Exploited Vulnerabilities catalog (Version 2026.06.11) [JSON data set]. https://github.com/cisagov/kev-data
8. Cybersecurity and Infrastructure Security Agency. (n.d.). Reducing the significant risk of known exploited vulnerabilities. Retrieved June 11, 2026, from https://www.cisa.gov/known-exploited-vulnerabilities-catalog/reducing-significant-risk-known-exploited-vulnerabilities
9. Dimitriadis, T., Gneiting, T., & Jordan, A. I. (2021). Stable reliability diagrams for probabilistic classifiers. Proceedings of the National Academy of Sciences of the United States of America, 118(8), Article e2016191118. https://doi.org/10.1073/pnas.2016191118
10. Forum of Incident Response and Security Teams. (2019). Common Vulnerability Scoring System version 3.1: Specification document. https://www.first.org/cvss/v3.1/specification-document
11. Forum of Incident Response and Security Teams. (2026a). Exploit Prediction Scoring System. Retrieved June 11, 2026, from https://www.first.org/epss/
12. Forum of Incident Response and Security Teams. (2026b). EPSS scores for June 11, 2026 [CSV data set]. https://www.first.org/epss/data_stats
13. Forum of Incident Response and Security Teams. (2026c). The EPSS model. Retrieved June 11, 2026, from https://www.first.org/epss/model
14. Forum of Incident Response and Security Teams. (2026d). Understanding EPSS probabilities and percentiles. Retrieved June 11, 2026, from https://www.first.org/epss/articles/prob_percentile_bins
15. Jacobs, J., Romanosky, S., Edwards, B., Adjerid, I., & Roytman, M. (2021). Exploit Prediction Scoring System. Digital Threats: Research and Practice, 2(3), Article 20. https://doi.org/10.1145/3436242
16. Jacobs, J., Romanosky, S., Suciu, O., Edwards, B., & Sarabi, A. (2023). Enhancing vulnerability prioritization: Data-driven exploit predictions with community-driven insights. In 2023 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW) (pp. 194–206). IEEE. https://doi.org/10.1109/EuroSPW59978.2023.00027
17. Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to information retrieval. Cambridge University Press. https://doi.org/10.1017/CBO9780511809071
18. Meyes, R., Lu, M., de Puiseau, C. W., & Meisen, T. (2019). Ablation studies in artificial neural networks. arXiv. https://doi.org/10.48550/arXiv.1901.08644
19. National Institute of Standards and Technology. (2026a). National Vulnerability Database. Retrieved June 11, 2026, from https://nvd.nist.gov/
20. National Institute of Standards and Technology. (2026b). National Vulnerability Database data feeds. Retrieved June 11, 2026, from https://nvd.nist.gov/vuln/data-feeds
21. National Institute of Standards and Technology. (2026c). NVD vulnerability APIs. Retrieved June 11, 2026, from https://nvd.nist.gov/developers/vulnerabilities
22. National Institute of Standards and Technology. (2026d). Understanding NVD vulnerability detail pages. Retrieved June 11, 2026, from https://nvd.nist.gov/vuln/vulnerability-detail-pages

23. National Institute of Standards and Technology. (2026e). Vulnerability metrics: Common Vulnerability Scoring System. Retrieved June 11, 2026, from https://nvd.nist.gov/vuln-metrics/cvss
24. National Institute of Standards and Technology. (2026f). NVD CVE modified feed: JSON 2.0 snapshot dated June 11, 2026 [Data set]. National Vulnerability Database.
25. Platt, J. C. (2000). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In A. J. Smola, P. L. Bartlett, B. Schölkopf, & D. Schuurmans (Eds.), Advances in large margin classifiers (pp. 61–74). MIT Press.
26. Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLOS ONE, 10(3), Article e0118432. https://doi.org/10.1371/journal.pone.0118432
27. Souppaya, M., & Scarfone, K. (2022). Guide to enterprise patch management planning: Preventive maintenance for technology (NIST Special Publication 800-40 Revision 4). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.SP.800-40r4
28. Sperandei, S. (2014). Understanding logistic regression analysis. Biochemia Medica, 24(1), 12–18. https://doi.org/10.11613/BM.2014.003