

CyberGuard AI: A Multi-Task Dual-Head SecBERT Architecture for Automated CVE Severity Triage and Threat Intelligence Generation

Aikagra Gupta

School of Computer Science, University of Petroleum and Energy Studies (UPES), Dehradun, Uttarakhand, India

Abstract - The rapid proliferation of software vulnerabilities catalogued in the National Vulnerability Database (NVD) creates an operational bottleneck for Security Operations Centers (SOCs), where thousands of Common Vulnerabilities and Exposures (CVEs) must be prioritized under strict time constraints. In this paper, we present CyberGuard AI, a multi-task dual-head neural architecture combining a cybersecurity-specialized transformer (SecBERT) with learned Common Weakness Enumeration (CWE) embeddings and live threat intelligence telemetry from CISA KEV and FIRST.org EPSS. Our architecture simultaneously executes two operational objectives in a single forward pass: (1) an actionable binary triage head discriminating severe threats ($CVSS \geq 7.0$) from standard risks ($CVSS < 7.0$), and (2) a granular four-class severity head classifying CVEs into LOW, MEDIUM, HIGH, and CRITICAL. Evaluated on a held-out test partition of 9,556 real-world NVD CVEs, the system achieves a binary triage ROC-AUC of 95.23%, an overall binary accuracy of 88.52%, a severe threat precision of 90.67%, a critical severity recall of 85.14%, and a top-2 tolerant classification accuracy of 95.81%. We report complete per-class metrics across all tiers, including minority classes, and integrate an automated generative advisory layer that synthesizes NIST SP 800-61 compliant response playbooks and exports structured OASIS STIX 2.1 bundles for enterprise SIEM ingestion.

Key Words: CVE Severity Classification, SecBERT, Multi-Task Learning, CWE Embeddings, Threat Intelligence, CVSS, Generative AI, STIX 2.1

1. INTRODUCTION

The Common Vulnerabilities and Exposures (CVE) catalogue, maintained by MITRE and scored by the National Vulnerability Database (NVD) using the Common Vulnerability Scoring System (CVSSv3), serves as the foundation for enterprise vulnerability management. With over 25,000 new vulnerabilities disclosed annually [1], Security Operations Centers (SOCs) face acute alert fatigue. Analysts must rapidly identify high-impact exposures requiring immediate mitigation while routing lower-risk items to standard patch maintenance windows. Under operational pressure, incident response teams prioritize answering a primary binary question: Does this disclosure represent an actionable severe threat requiring emergency patching? Manual triage is hindered by subjective scoring variations, where independent

evaluators often diverge on borderline ratings such as CVSS 6.9 (MEDIUM) versus 7.1 (HIGH) [2]. Furthermore, general-domain language models lack grounding in low-level vulnerability concepts such as kernel pointer dereferencing, heap corruption primitives, and protocol exploitation sequences.

To address these challenges, we present CyberGuard AI, built upon four technical contributions:

1. A joint multi-task neural architecture that simultaneously outputs binary actionable triage and four-class granular severity from a shared domain-adapted transformer backbone in a single forward pass.
2. An 800-dimensional unified feature representation combining 768-dimensional contextual SecBERT embeddings with learned 32-dimensional CWE vector representations across 307 weakness types.
3. Real-time threat intelligence enrichment prefixing input advisories with CISA Known Exploited Vulnerabilities (KEV) flags and FIRST.org Exploit Prediction Scoring System (EPSS) probabilities.
4. A production-grade inference engine coupled with a generative threat response agent that produces NIST SP 800-61 playbooks and standardized OASIS STIX 2.1 bundles.

2. RELATED WORK

2.1 Vulnerability Scoring Automation

Early methods for automated CVSS prediction utilized term frequency-inverse document frequency (TF-IDF) feature extraction with Support Vector Machines (SVM) or Random Forests [3]. While computationally efficient, these models fail to capture deep semantic dependencies or distinguish nuanced impact descriptions, such as localized resource exhaustion versus remote code execution.

2.2 Transformers in Security NLP

General transformer architectures [4] significantly advanced technical text processing. Shahid and Debar [5] developed CVSS-BERT to predict CVSS metric vector components with gradient-based saliency maps for interpretability. Das et al. [6] introduced V2W-BERT, demonstrating hierarchical multiclass classification of software vulnerability descriptions to CWE categories.

However, general-domain backbones encounter out-of-vocabulary limitations on specialized security tokens.

2.3 Domain Pre-Training and Multi-Task Architectures

Cybersecurity-adapted models such as SecBERT [7] and SecureBERT [8] address vocabulary mismatches by pre-training directly on vulnerability corpora, advisories, and technical threat bulletins. Multi-task learning (MTL) provides an effective paradigm for training shared representations across complementary tasks [9]. CyberGuard AI leverages SecBERT within an MTL formulation to concurrently address high-level triage and granular severity assessment.

3. DATASET AND PREPROCESSING

3.1 Enriched Vulnerability Corpus

Our dataset is compiled from NIST NVD CVSSv3 records, augmented with two live threat intelligence sources:

CISA Known Exploited Vulnerabilities (KEV) [10]: Ground-truth indicators of confirmed in-the-wild adversary exploitation.

FIRST.org EPSS [11]: Calibrated probabilities estimating the empirical likelihood of exploitation within 30 days. Records with missing descriptions or CVSSv3 base vectors were excluded, yielding 95,559 fully labeled CVE instances.

Table - 1: Class Distribution of Enriched CVE Corpus

Severity Class	CVSSv3 Base	Record Count	%
CRITICAL	9.0 – 10.0	14,198	14.86
HIGH	7.0 – 8.9	41,634	43.57
MEDIUM	4.0 – 6.9	37,912	39.68
LOW	0.1 – 3.9	1,815	1.90
Total		95,559	100.0

Binary triage labels follow the industry standard: records with CVSS ≥ 7.0 are designated Severe Actionable (55,832 samples, 58.4%), while records below 7.0 are designated Standard Risk (39,727 samples, 41.6%).

3.2 Stratification and Prompt Enrichment

The dataset was split using stratified sampling into 85% training (81,225 samples), 5% validation (4,778 samples), and 10% strictly held-out test data (9,556 samples).

Each vulnerability narrative is converted into a structured prompt prior to tokenization:

[Threat Context: Standard Advisory | Exploit Probability: 14.2% | Weakness: CWE-79] <raw description>

When a vulnerability is flagged in CISA KEV, the prefix dynamically updates to [Threat Context: ACTIVELY EXPLOITED IN THE WILD].

4. SYSTEM ARCHITECTURE

4.1 Backbone and CWE Embedding Fusion

The representation encoder initializes from Jackduma/SecBERT, a 12-layer transformer with 110M parameters. Token-level hidden states are mean-pooled across sequence length $L \leq 256$:

$$h_{\text{text}} = (1/L) \sum H_i \in \mathbb{R}^{768} \quad (1)$$

Concurrently, the vulnerability's weakness identifier is mapped through an embedding matrix covering 307 CWE classes:

$$e_{\text{cwe}} = E_{\text{CWE}}(\text{idx}) \in \mathbb{R}^{32} \quad (2)$$

The two representations are concatenated into an 800-dimensional fused feature vector:

$$f = [h_{\text{text}} \parallel e_{\text{cwe}}] \in \mathbb{R}^{800} \quad (3)$$

4.2 Dual Output Heads and Loss Formulation

The fused vector f passes through a dense projection layer with ReLU activation and 20% dropout:

$$z = \text{Dropout}(\text{ReLU}(W_{\text{proj}} \cdot f + b_{\text{proj}})) \in \mathbb{R}^{256} \quad (4)$$

The latent representation z feeds into two parallel linear heads:

$$\hat{y}_{\text{4class}} = \text{Softmax}(W_1 z + b_1) \in \mathbb{R}^4 \quad (5)$$

$$\hat{y}_{\text{binary}} = \text{Softmax}(W_2 z + b_2) \in \mathbb{R}^2 \quad (6)$$

The overall objective is optimized jointly using balanced cross-entropy losses:

$$L_{\text{total}} = 0.5 \cdot L_{\text{4class}} + 0.5 \cdot L_{\text{binary}} \quad (7)$$

This joint structure provides two actionable outputs in a single forward pass, enabling simultaneous high-level triage routing and granular severity scoring.

5. EXPERIMENTAL RESULTS

5.1 Training Setup

Fine-tuning was conducted on an NVIDIA RTX 5050 GPU (8 GB VRAM) using BF16 mixed-precision acceleration.

We trained with AdamW (learning rate 2×10^{-5} , weight decay 0.01), linear warmup with cosine decay, batch size 32, and sequence length 256 over 3 epochs.

5.2 Verified Benchmark Performance

All reported performance metrics are evaluated strictly on the 9,556 held-out test samples.

Table - 2: CyberGuard AI Test Set Benchmarks (n = 9,556)

Evaluation Metric	Head	Test Score
Binary Triage ROC-AUC	Head 2	95.23%
Top-2 Tolerant Accuracy	Head 1	95.81%
Severe Threat Precision	Head 2	90.67%
Severe Threat Recall	Head 2	89.58%
Binary Triage Accuracy	Head 2	88.52%
Binary Triage Weighted F1	Head 2	88.53%
Critical Severity Recall	Head 1	85.14%
Granular 4-Class Accuracy	Head 1	78.34%
Granular 4-Class Macro F1	Head 1	70.10%
Granular 4-Class Weighted F1	Head 1	78.65%

5.3 Per-Class Evaluation (Including Minority Classes)

Table 3 details the performance across all individual severity classes on the held-out test partition. The test breakdown demonstrates that the model does not collapse on the heavily imbalanced LOW class (1.9% support), identifying 106 of 181 LOW instances (58.56% recall). For CRITICAL threats, the model achieves 85.14% recall (1,209 of 1,420 critical vulnerabilities identified), ensuring minimal false negatives for high-severity exposures.

In binary triage, the model correctly identifies 5,002 of 5,584 Severe Actionable instances (89.58% recall, 90.67% precision) with 515 false positives and 582 false negatives out of 9,556 cases.

Table - 3: Detailed Per-Class Severity Metrics (n = 9,556)

Class	Support	Precision	Recall	F1-Score
LOW	181	38.27%	58.56%	46.29%
MEDIUM	3,791	84.53%	81.56%	83.02%
HIGH	4,164	82.46%	73.94%	77.97%
CRITICAL	1,420	64.07%	85.14%	73.12%
Macro Avg	9,556	67.33%	74.80%	70.10%
Weighted Avg	9,556	79.71%	78.34%	78.65%

5.4 Architectural Progression

Table - 4: Architectural Development Progression

Iter.	Architecture	4-Class Acc.	Triage AUC
v1-v3	TF-IDF + LR / SVM	68.20%	–
v5	DistilBERT (General)	74.50%	–
v6	DeBERTa-v3 + Focal Loss [12]	77.20%	–
v8	RoBERTa + CWE [13]	78.90%	–
v11	SecBERT + KEV/EPSS Telemetry	80.50%	–
v12	SecBERT + CWE + Dual-Head	78.34%	95.23%

While v12 trades approximately 2% in four-class accuracy relative to the single-head v11 baseline, it provides a high-confidence operational triage head (95.23% ROC-AUC) and 95.81% top-2 tolerant accuracy.

6. DEPLOYMENT AND GENERATIVE AGENT

6.1 Inference Service and Dashboard

CyberGuard AI is deployed via a FastAPI REST service (/api/v1/classify) and an interactive Streamlit dashboard. End-to-end inference latency is under 10 ms on GPU and under 50 ms on modern CPU hardware. The public cloud demo operates in hybrid mode: model weights are omitted due to repository size constraints, and classification falls back to Gemini-based inference.

6.2 Generative Advisory and STIX 2.1 Serialization

Model outputs feed an automated generative agent (Google Gemini 2.5 Flash, with Groq and Ollama fallback) that synthesizes incident response playbooks aligned with NIST SP 800-61 Rev. 2, maps vulnerabilities to CIS Controls and ISO 27001, and exports validated OASIS STIX 2.1 JSON packages [14] for SIEM platforms including Splunk and Microsoft Sentinel.

7. CONCLUSIONS

This paper presented CyberGuard AI, a multi-task dual-head SecBERT framework for vulnerability severity prediction and triage. Incorporating domain-specific pretraining, CWE weakness embeddings, and live threat telemetry within a joint classification objective, the model achieves 95.23% ROC-AUC on binary triage, 95.81% top-2 tolerant accuracy, and 85.14% recall on critical exposures across 9,556 held-out test samples. Complete per-class

reporting confirms balanced detection across all severity levels.

Future work will investigate continuous online learning pipelines to ingest newly published CVE feeds, multilingual advisory analysis, and graph neural network modeling of multi-hop vulnerability exploitation chains.

8. REFERENCES

- 1) National Institute of Standards and Technology, "National Vulnerability Database (NVD) Statistics Dashboard," 2024. [Online]. Available: <https://nvd.nist.gov/vuln/search/statistics>
- 2) L. Allodi and F. Massacci, "Comparing vulnerability severity and exploits using case-control studies," *ACM Trans. Inf. Syst. Secur.*, vol. 17, no. 1, pp. 1–20, 2014.
- 3) P. Holm, M. Ekstedt, and D. Andersson, "Empirical analysis of system-level vulnerability metrics through text mining," *IEEE Trans. Dependable Secure Comput.*, vol. 9, no. 6, pp. 887–900, 2012.
- 4) J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, Minneapolis, MN, pp. 4171–4186, 2019.
- 5) M. R. Shahid and H. Debar, "CVSS-BERT: Explainable natural language processing to determine the severity of a computer security vulnerability from its description," in *Proc. 20th IEEE ICMLA*, Pasadena, CA, USA, pp. 1600–1607, Dec. 2021, doi: 10.1109/ICMLA52953.2021.00256.
- 6) A. Das, D. Serra, M. Halappanavar, A. Pothen, and E. Al-Shaer, "V2W-BERT: A framework for effective hierarchical multiclass classification of software vulnerabilities," in *Proc. 2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA)*, Porto, Portugal, pp. 1–10, Oct. 2021, doi: 10.1109/DSAA54385.2021.00018.
- 7) Jackaduma, "SecBERT: A pre-trained language model for cyber security text," *HuggingFace Model Hub*, 2021. [Online]. Available: <https://huggingface.co/jackaduma/SecBERT>
- 8) E. Aghaei, X. Niu, W. Shadid, and E. Al-Shaer, "SecureBERT: A domain-specific language model for cybersecurity," in *Proc. 18th EAI SecureComm*, Kansas City, MO, USA, 2022. [Online]. Available: <https://arxiv.org/abs/2204.02685>
- 9) R. Caruana, "Multitask learning," *Machine Learning*, vol. 28, no. 1, pp. 41–75, 1997.
- 10) Cybersecurity and Infrastructure Security Agency, "Known Exploited Vulnerabilities Catalog," 2024. [Online]. Available: <https://www.cisa.gov/known-exploited-vulnerabilities-catalog>
- 11) FIRST.org, "Exploit Prediction Scoring System (EPSS)," 2024. [Online]. Available: <https://www.first.org/epss/>
- 12) T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE ICCV*, pp. 2980–2988, 2017.
- 13) Y. Liu et al., "RoBERTa: A robustly optimized BERT pretraining approach," *arXiv:1907.11692*, 2019.
- 14) OASIS Open, "STIX Version 2.1," OASIS Standard, 2021. [Online]. Available: <https://oasis-open.github.io/cti-documentation/stix/intro>

9. BIOGRAPHY

Aikagra Gupta is an undergraduate scholar in the School of Computer Science at the University of Petroleum and Energy Studies (UPES), Dehradun, India. His research interests include cybersecurity threat intelligence, applied deep learning, natural language processing for software vulnerability assessment, and autonomous security automation.