UNIVERSITI TEKNOLOGI MARA

CSP760

PREDICTING VULNERABILITY SUSCEPTIBILITY IN MALAYSIAN BANK USING SUPERVISED MACHINE LEARNING

STUDENT NOR ADANI BINTI KAMAL MOHAMAD NASIR (2024782087)

SUPERVISOR DR SITI ARPAH BINTI AHMAD

Date: 2 November 2025

UNIVERSITI TEKNOLOGI MARA

TABLE OF CONTENT

(ADANI,2025)

SUMMARY PROJECT

Problem	Research Questions	Hypothesis	Objectives	Scope	Significance
Malaysian banks rely heavily on **Vulnerability Assessment** (VA) tools such as Nessus, Qualys, and OpenVAS that only detect *existing* **vulnerabilities** but cannot *predict future threats*. This leaves financial institutions unprepared for advanced cyberattacks.	**RQ1:** What are the limitations of current **vulnerability assessment** tools in predicting future attack trends? **RQ2:** To what extent can machine learning models improve **vulnerability** prioritization and remediation by forecasting attack patterns? **RQ3:** How effective is the proposed ML model compared to traditional VA methods in predicting **vulnerability** susceptibility?	**H1:** Machine Learning algorithms can enhance the predictive capabilities of **vulnerability assessments**. **H2:** Random Forest (RF) and Neural Network Regression (NNR) are suitable models for forecasting **vulnerability susceptibility**.	1. To design and propose a machine learning model for predicting cyberattack susceptibility using **vulnerability Assessment** data. 2. To develop and implement a supervised ML-based predictive system using Kaggle and Tenable datasets. 3. To evaluate the accuracy and performance of the model in forecasting potential vulnerabilities.	- **Data:** **Vulnerability** from Kaggle, CVE vulnerability, exploitDB, Tenable datasets. - **Tools:** Orange Data Mining, Python (Jupyter Notebook). - Models: Random Forest (classification), Neural Network Regression (regression). - Evaluation: Accuracy, Precision, Recall, F1-score, MAE, RMSE.	The study enhances the cybersecurity posture of Malaysian banks by shifting from reactive to **predictive vulnerability** management. It supports AI-driven, enables faster remediation prioritization.

UPDATE WEEK 4

No	Item	Key processing done (this week)	Status / deliverable
1	**Kaggle** — 1,314 rows (vulnerability dataset)	Mapped CVSS → Severity column, mend column names, removed exact duplicates.	✅ Done (kaggle.csv)
2	**NVD** — 41,241 rows (filtered to 2025).	Extracted CVE IDs, CVSS, published dates, removed null CVE rows. Need to request the API key.	✅ Done (cve_2025_clean.csv)
3	**ExploitDB** — 46,922 rows.	Scraped page by page to extracted exploit data. Target info is titles, type attack, date, added port/platform mapping and removed corrupted rows	✅ Done (exploitdb_extracted_date.csv)
4	**Tenable (public plugins)** — 20,000 rows.	Extracted Plugin ID, Title, CVSS, Family, Solution; aligned Plugin ID, CVE. Downloadling took a lot of time. If cut off, need to restart	⚙️ In progress
5	**Merge & clean**	Merged all sources → merged_vulnerability_data.csv removed null/empty rows, added **key column** (source, family, asset, Label) and Severity (Low / Medium / High / Critical)	⚙️ Merging & cleaning ongoing
6	**Jupyter scripts & snippets**	Scripts for: (1) fetch NVD, ExploitDB, Tenable (2) Data merging & cleaning (3) Column add/removed depends on key column	All related have been document
7	**Next steps**	Feature engineering (Family, Asset Type), EDA charts, prepare subset for Orange (RF + NNR)	In Planned

DATASET SOURCES

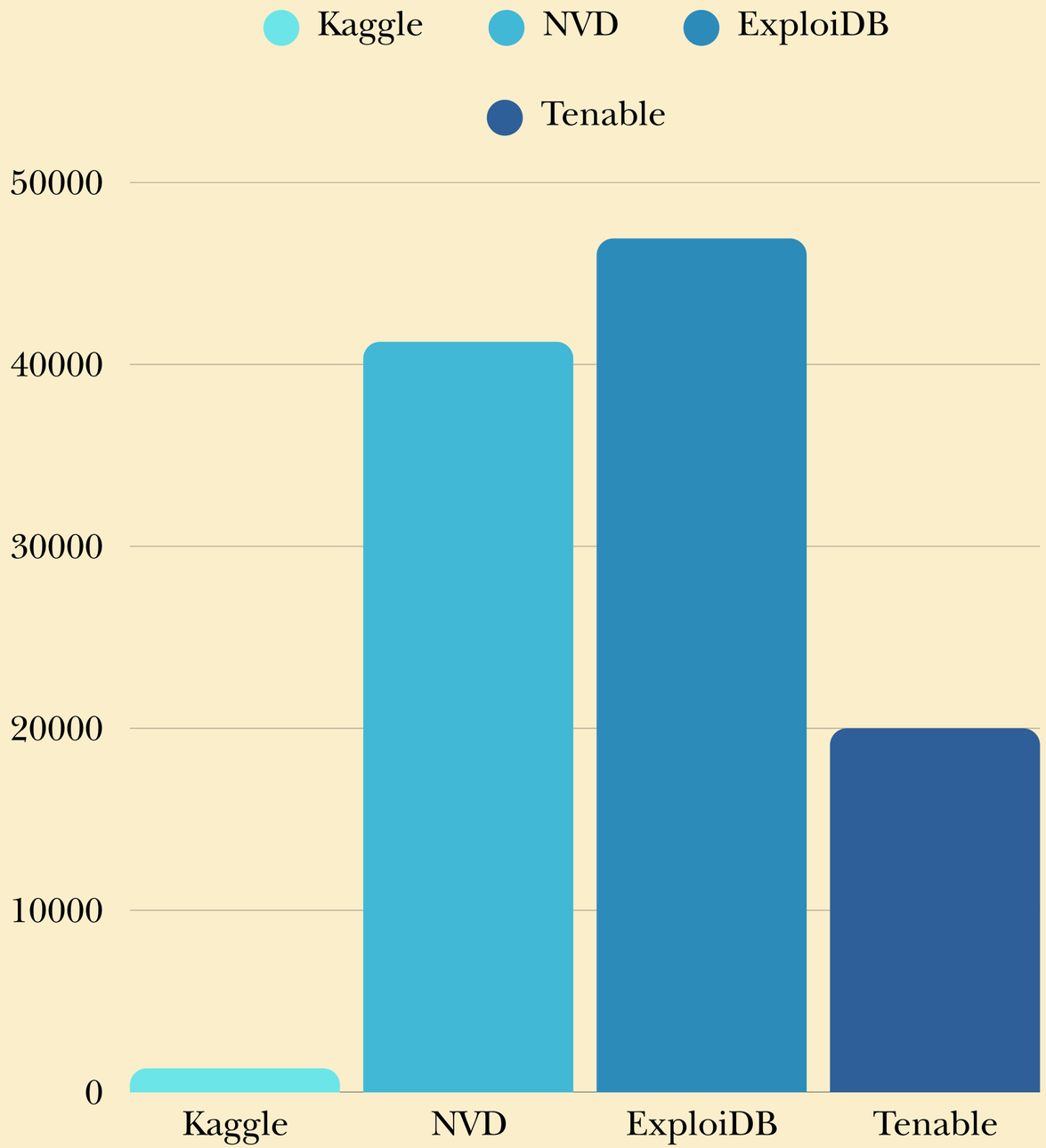
Source: Publicly-available data (CVE, ExploitDB, Kaggle, Tenable) used to build a proxy dataset.

Reason: Original bank dataset is sensitive and privacy. I created a proxy that mirrors a real vulnerability data.

New **columns** added for analysis: Source, Family, Asset Type, Bank (Label - Y/N)

Dataset: 100 000 row (80 000)

Kaggle (1,314), NVD/CVE 2025 (41,241), ExploitDB (46,922), Tenable (20,000)



vulnerability_assessment_data.csv

2 | Code | Download

Data Card | Code (1) | Discussion (0) | Suggestions (0)

vulnerability_assessment_data.csv (12.41 kB)

Detail | Compact | Column | 10 of 10 columns

About this file

Suggest Edits

The dataset used in this project is provided in a CSV (Comma Separated Values) file format. The CSV file contains structured data related to vulnerability assessments, encompassing various attributes such as vulnerability descriptions, severity levels, vulnerability types, affected software, discovery dates, patch dates, risk scores, description lengths, and days to patch. Each row in the CSV file represents a distinct vulnerability instance, providing a comprehensive overview of security vulnerabilities across different software systems. The CSV format allows for easy access, manipulation, and analysis of the data using programming languages such as Python and data analysis tools like pandas

ID	Description	Severity	Vulnerability Type	Affected Softwa
Unique identifier for each vulnerability instance.	Severity: Level of severity associated with the vulnerability (e.g., Low, Medium, High).	Vulnerability Type: Type or category of vulnerability (e.g., XSS, CSRF, Buffer Overflow).	Affected Software: Software or system affected by the vulnerability.	Discovered Date: D and time when the vulnerability was discovered.

Discovered Date	Patch Date	# Risk Score	# Description Length	# Days to Patch
Discovered Date: Date and time when the vulnerability was discovered.	Patch Date: Date and time when the vulnerability was patched or fixed.	Risk Score: Numerical score representing the risk associated with the vulnerability.	Description Length: Length of the vulnerability description text.	Days to Patch: Numb of days taken to patc the vulnerability after discovery.

Data Explorer

Version 1 (12.41 kB)

vulnerability_assessment_data

Summary

1 file

10 columns

DATASET

Title	Total Dataset Used	What Dataset/Samples	Train/Test Split
CVE Severity Prediction From Vulnerability Description – A Deep Learning Approach (2024)	7,765 CVEs (test set); total dataset size not fully stated	NVD (CVE descriptions, CVSS severity levels)	Not explicitly stated; likely standard 80/20 based on NLP tasks
A Vulnerability Severity Prediction Method Based on Bimodal Data and Multi-Task Learning (2023)	~20,000+ samples	Combined textual (NVD) + numerical (CVSS metrics)	80% training / 20% testing
An Improved Vulnerability Exploitation Prediction Model with Novel Cost Function and Custom Trained Word Vector Embedding (2021)	146,183 CVEs	NVD + ExploitDB + Symantec attack signatures	Train/test split + 10-fold cross-validation
Improving Vulnerability Prediction of JavaScript Functions Using Process Metrics (2021)	12,125 functions	SATE IV dataset + GitHub functions labeled vulnerable/non-vulnerable	80% train / 10% validation / 10% test
Common Vulnerability Scoring System Prediction based on Open Source Intelligence Information Sources (2022)	88,979 CVEs (2016–2021)	NVD descriptions + web-scraped OSINT (blogs, vendor advisories, GitHub)	Split not stated; focused on large-scale DL-based analysis
Analysis and Prediction of Cyber-Threat Fragments in Banking Critical Infrastructure Sector (2024)	Approx. 1,500 banking-related incidents	Real cyber threat data from banking sector. incident logs, security feeds, and open-source threat repositories	Not stated
Predicting Vulnerability Susceptibility in Malaysian Banks Using Supervised Machine Learning	Kaggle (1,314), NVD/CVE 2025 (41,241), ExploitDB (46,922), Tenable (20,000) = 109477 (80k)	Merged datasets (Kaggle + Tenable + CVE, ExploitDB) with labelling bank related	Planned 70/30 split

DATA COLLECTION → KAGGLE, NVD, EXPLOIT DB, TENABLE

IDENTIFY KEY FEATURES → PLUGIN ID, VULNERABILITY NAME, CVSS SCORE, SEVERITY, CVE, APPLICATION

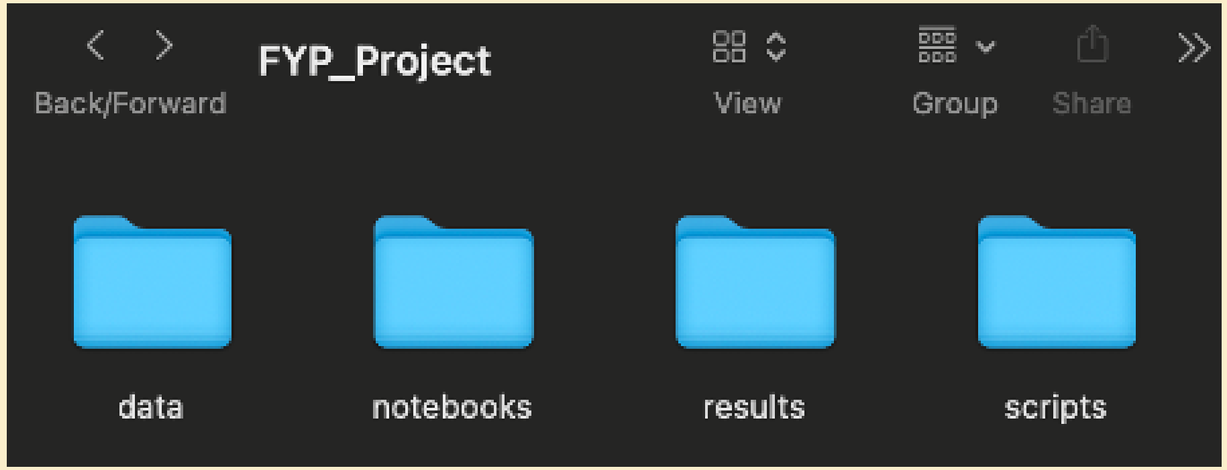
REMOVE USELESS COLUMN

DERIVE SEVERITY (FROM CVSS SCORE)

REMOVE DUPLICATES → MERGING ALL SOURCES

CLEAR NULLS

FEATURE ENGINEER ← ASSET TYPE, FAMILY, BANK RELATED (LABEL)



/ Documents / FYP_Project / scripts /

Name	Modified	File Size
cve.ipynb	40 minutes ago	24.8 KB
ExploitDB.ipynb	38 minutes ago	29.4 KB
compare&combine.ipynb	8 minutes ago	32.7 KB
tenable.ipynb	51 seconds ago	10 KB

Rating	CVSS Score
None	0.0
Low	0.1-3.9
Medium	4.0-6.9
High	7.0-8.9
Critical	9.0-10.0

DATA PROCESSING PIPELINE

CSP760

THANK YOU