

Supplementary Document: Bibliometric Data Pipeline for Insider Threat Detection (SLR)

1. Overview

This document outlines the methodology used to generate the bibliometric visualizations for the Systematic Literature Review (SLR) titled "Insider Threat Detection Using Machine Learning." To ensure academic strength and 100% representation of the 82-paper research corpus, the process utilized a data pipeline. The pipeline combined data extraction capabilities of NotebookLM with the quantitative visualization powers of Python (Google Colab) and MS Excel. This approach ensures that every figure ranging from keyword frequency treemaps and word clouds to publisher distributions and JCR quartile rankings is accurate. To maintain data integrity across the 82 selected papers, a dual-stream data pipeline was utilized:

Stream A (Content-Based): Qualitative thematic extraction using NotebookLM and high-resolution visualization via Python (Google Colab).

Stream B (Structural-Based): Quantitative metadata mapping using Zotero and MS Excel. This was also manually validated against the Web of Science (WoS) JCR database.

This dual approach ensures that every figure is grounded in manually audited metadata, providing a transparent and reproducible data for the study's bibliometric analysis.

2. Part I: Content-Based Bibliometric Analysis (Thematic)

This phase transformed the unstructured text of 82 full-text PDFs into structured thematic data.

2.1. AI-Driven Frequency Audit (NotebookLM)

All 82 PDFs were uploaded into Google **NotebookLM** for an analysis of leading or most frequent keywords. We utilized a comprehensive audit prompt as shown below:

The Prompt:

The analysis was executed using specific instructions to ensure statistical accuracy:

Prompt Snippet: *"Perform an exhaustive Frequency Audit of all 82 papers... Create a Table representing the global dataset. If a keyword appears in 40 papers, we expect to see the number 40... identify 'Minority Themes' found in outlier papers to ensure 100% representation."*

Statistical Calibration:

To ensure the data was ready for visualization, two specific metrics were calculated:

1. **Coverage % (Prevalence):** $(f / 82) \times 100$. This accounts for papers belonging to multiple categories (e.g., a paper using both *Deep Learning* and *Anomaly Detection*).
2. **Relative % (Distribution):** $(f / \sum f) \times 100$. This was the metric used for the Treemap and Word Cloud graphs.

2.2 Python Visualization Logic (Figures 2 & 3)

The audit data was moved to **Google Colab** to generate Figures 2 and 3.

- **Figure 2 (Treemap):** Size was mapped to Relative % using a customized colors for visual distinction between the keywords.
- **Figure 3 (Word Cloud):** Configured with `collocations=False` to preserve technical phrases (e.g., "Deep Learning") as single units rather than fragmented words.

3. Technical Implementation: Python Source Code

The following scripts were executed to ensure high-resolution (300DPI) output suitable for publication.

Code for Figure 2: Top 10 Keyword Treemap

```
import plotly.express as px
import pandas as pd
import textwrap
# 1. Dataset from NotebookLM Audit
data = {
    'Keyword': ['Anomaly Detection', 'Deep Learning', 'Behavioral Analysis', 'Graph Neural Networks',
                'Ensemble Learning', 'Data Imbalance Handling', 'Federated Learning',
                'NLP / Sentiment Analysis', 'Explainable AI (XAI)', 'IoT Security'],
    'Frequency': [42, 40, 24, 21, 19, 17, 12, 9, 8, 7],
    'Relative_%': [21.1, 20.1, 12.1, 10.6, 9.6, 8.5, 6.0, 4.5, 4.0, 3.5]
}
df = pd.DataFrame(data)

# 2. Styling and Rendering
def wrap_text(text, width=12): return "<br>".join(textwrap.wrap(text, width=width))
df['Label'] = "<b>" + df['Keyword'].apply(wrap_text) + "</b><br>" + df['Frequency'].astype(str) + " (" + df['Relative_%'].astype(str) + "%)"

My_colors = ['#8dd3c7', '#ffffb3', '#bebada', '#fb8072', '#80b1d3', '#fdb462', '#b3de69', '#fccde5', '#d9d9d9', '#bc80bd']
fig = px.treemap(df, path=['Label'], values='Relative_%', color='Keyword', color_discrete_sequence=My_colors)
fig.update_layout(title_text='<b>Figure 2. Top 10 Keyword Treemap Analysis (N=82)</b>', title_x=0.5)
fig.show()
```

Code for Figure 3: Thematic Word Cloud

```
from wordcloud import WordCloud
import matplotlib.pyplot as plt

# 1. Thematic Dictionary (Top 10 + Minority Themes)
keywords_freq = {
    'Anomaly Detection': 42, 'Deep Learning': 40, 'Behavioral Analysis': 24,
    'Graph Neural Networks': 21, 'Ensemble Learning': 19, 'Data Imbalance Handling': 17,
    'Federated Learning': 12, 'NLP / Sentiment Analysis': 9, 'Explainable AI (XAI)': 8,
    'IoT Security': 7, 'Isolation Forest': 14, 'XGBoost': 13, 'Support Vector Machine': 12,
    'Hidden Markov Models': 11, 'Word2Vec': 9, 'Autoencoders': 9, 'Zero Trust': 6
}

# 2. Configure and Generate Word Cloud
wc = WordCloud(width=2000, height=1000, background_color='white', colormap='plasma',
               collocations=False, max_words=100, random_state=42).generate_from_frequencies(keywords_freq)

# 3. Render Visualization
plt.figure(figsize=(20, 10))
plt.imshow(wc, interpolation='bilinear')
plt.axis("off")
plt.title("Figure 3. Word Cloud of Insider Threat Detection Keywords", fontsize=30, pad=20)
plt.show()
```

The screenshot illustrates the NotebookLM interface during the Deep Audit process. On the left, a list of 82 research papers is shown, each with a small thumbnail and a brief title. In the center, a table provides a quantitative overview of the keyword frequency and calibration across ten categories. The right side features a 'Studio' workspace where users can generate various types of visualizations based on the ingested data.

Figure 1. Interface of the Deep Audit Process in NotebookLM.

The above snapshot demonstrates the ingestion of the 82-paper research corpus (left pane) and the generation of the Keyword Frequency and Treemap Calibration Table (center pane) used as the primary data source for subsequent Python-based visualizations.

4. Part II: Structural Bibliometric Analysis (Metadata)

This phase utilized the bibliographic metadata of the 82 papers to map the research landscape's quality and trends.

4.1. Metadata Extraction and Data Cleaning (Zotero)

The primary metadata was managed via **Zotero**. To ensure 100% accuracy, a "Clean-Room" approach was adopted:

- Export:** A CSV file was generated from the Zotero library containing the 82 studies.
- Manual Verification:** Every entry was manually cross-referenced against the original PDF metadata to correct any errors in **Publication Year**, **Journal Name**, and **Publisher**.

4.2. Quantitative Mapping via MS Excel (Figures 4–7)

Using the verified CSV, MS Excel was utilized to generate the structural visualizations:

- Figure 4 (Publication by Year):** A column chart illustrating publication volume between 2021–2025.
- Figure 5 & 6 (Publisher/Journal Distribution):** Mapping the corpus to major academic houses (IEEE, Elsevier, Springer).
- Figure 7 (JCR Quality Audit):** Each journal was manually cross-referenced with the **Web of Science (WoS)** database to identify its **JCR Quartile (Q1–Q4)**.

5. Summary of the Data Pipeline

The following data served as the input for the final visualizations:

Stage	Process	Primary Tool	Output
Collection	Search & Ingestion	Zotero	Master Bibliography
Thematic Audit	Content Deep-Scan	NotebookLM	Frequency Table
Thematic Viz	Visual Generation	Python (Colab)	Figure 2 & 3
Metadata Audit	Validation	Manual Review	Verified CSV
Structural Viz	Trend Mapping	MS Excel	Figure 4, 5, & 6
Quality Audit	JCR Ranking	Web of Science MS Excel	Figure 7 (Bar Chart)