Generative AI-Driven Data Transformation for Enhanced Machine Learning Performance (Student Abstract)

Christopher MacDowell, Sarah Setiawan, Carol Jim, Ahmed Salem

Department of Computer Science and Information Technology
Hood College, Frederick, Maryland 21701
{crm12, sss13, jim, salem}@hood.edu

Abstract

Machine Learning (ML) models have significant potential across research and industry to enable data-driven insights and decision-making. Their performance relies on input data quality, but real-world datasets often contain imperfections, making data preprocessing essential yet time-consuming. Our research proposes a proof-of-concept model using Generative Artificial Intelligence (GenAI) to analyze and transform data for supervised ML classification. The results from the GenAI models will be compared with traditionally preprocessed data to evaluate effectiveness. Preliminary results indicate that incorporating GenAI models into the preprocessing pipeline show potential in improving ML's classification performance.

Code — https://github.com/macdowellcr/AAAI_Student_Abstract

Dataset 1 — https://research.unsw.edu.au/projects/unsw-nb15-dataset

Dataset 2 — https://ieee-dataport.org/documents/real-time-dataset-idsiot2024

Introduction

Approximately 80% of a data scientist's time is spent on data preprocessing tasks, centered on cleaning and transformation (Cote et. al. 2024). Data cleaning involves rectifying errors and inconsistencies, while transformation includes scaling, normalization, and feature engineering. (Dhawas et. al. 2024). In the context of preprocessing a dataset for ML application, correct transformation is of paramount importance. Standard frameworks for data transformation are exceedingly human-centric, requiring the data scientist to manually identify and manipulate the dataset to achieve acceptable results. This is especially true in cybersecurity, where network traffic datasets are often large, high-dimensional, and imbalanced, making manual preprocessing a significant challenge. Our research proposes using Generative Artificial Intelligence (GenAI) to automate and improve the data transformation process for ML-based classification of network traffic. We test this by comparing the performance of ML models on datasets preprocessed by GenAI with those preprocessed using traditional manual methods. We utilize the well-known UNSW-NB15 dataset (Moustafa, Slay 2015a) and the newer, more complex IDSIoT2024 dataset (Koppula, Leo Joseph 2025) to evaluate our approach. Our preliminary results show that GenAI-driven preprocessing can significantly enhance ML classification performance, particularly on noisy, real-world data.

Methodology

Baseline Dataset Preparation

Two datasets were used for our experiments. The first, UNSW-NB15, is composed of synthesized raw traffic packets simulating both normal and malicious network activity in nine attack categories. The dataset has been cleaned by its authors but has not been transformed for ML application (Moustafa, Slay 2015b). For our baseline, we performed minimal manual preprocessing, which involved dropping three categorical variables and applying Z-score normalization to forty numerical variables. The second dataset, IDSIoT2024, is a more recent and complex real-world dataset for IoT cyber-attack detection. This dataset is considered "dirty" due to its raw, untransformed state, reflecting the challenges of real-world data. The manual preprocessing consisted of first, splitting the data into train and test sets before cleaning to prevent data leakage. Then, the same cleaning steps were applied to both train and test sets separately to ensure data consistency. Once processed, both UNSW-NB15 and IDSIoT24 were used to train and test Decision Tree (DT), Random Forest (RF), K-Nearest Neighbor (KNN), and Gaussian Naive Bayes (GNB) algorithms. The resulting classification metrics from these baseline models were used for comparison against the GenAI-preprocessed datasets.

GenAI Experimental Runs

Three of the latest GenAI models, Gemini 2.5 Pro, DeepSeek V3.1 and GPT-5, were programmed to perform data transformation on the original dataset. Figure 1 depicts the steps in this process. The process entails submitting a raw network traffic dataset to a GenAI reasoning model through a structured API. A Python script orchestrates this via a chained prompt, instructing the GenAI LLM. The initial instruction is to complete a comprehensive exploratory data analysis (EDA) on the dataset. Following the EDA, the

LLM is tasked with generating a robust data preprocessing Python script. The primary output of this phase is a validated Python preprocessing script. The same set of instructions were programmed into the three GenAI models via Application Program Interface (API) calls. If the cleaning code generated by the LLM resulted in errors, the chained prompt is adjusted to address the errors. If the data transformation code generated and returned by the programmed GenAI models compiled without error, it was run through the same Decision Tree, Random Forest, and K-Nearest Neighbor algorithms as the baseline dataset. Results from these models were then compared against the baseline and one another.

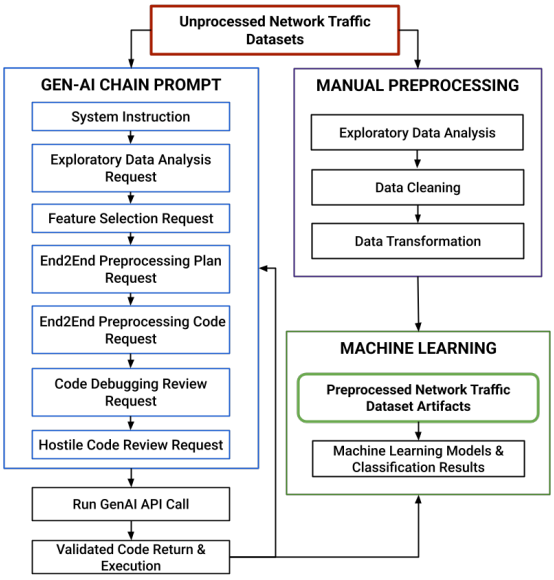

Figure 1: Experimental Process Flowchart

Results and Discussion

It is notable that the GenAI models generally did not perform as well as the Manual method on the UNSW-NB15 dataset, where the Manual preprocessing consistently yielded the higher scores. For example, shown in Table 1: the Manual RF F-1 Score was 0.62, while the GenAI models only reached 0.57 to 0.58 which highlights that all methods struggled with the dataset's class imbalance.

The performance comparison on the IDS-IoT2024 dataset strongly highlights the efficacy of the GenAI models' generated preprocessing code. The GenAI models consistently matched or outperformed the Manual approach across nearly all algorithms and metrics on IDS-IoT2024. For example, in Table 2, the robust KNN, DT, and RF algorithms, the GenAI models achieved high Accuracy scores, generally around 0.95, significantly exceeding the Manual scores that ranged from 0.81 to 0.88. This superior performance extended to complex metrics like the F-1 Score, where the GenAI models scored 0.87 to 0.91 for RF, compared to the Manual score of 0.73. Furthermore, on the challenging GNB algorithm, the GenAI models showed substantial improvement, yielding F-1 Scores between 0.64 and 0.79 compared to the manual score of .33 which demonstrates their ability to prepare the IDS-IoT2024 data for various types of classifiers compared to the human-driven Manual process.

UNSW-NB15 (Macro)				
ML Metric	**Manual**	**DeepSeek V3**	**Gemini 2.5 Pro**	**GPT-5**
K-Nearest Neighbor				
Accuracy	0.83	0.75	0.75	0.73
Precision	0.52	0.45	0.46	0.44
Recall	0.64	0.53	0.57	0.55
F-1 Score	0.54	0.47	0.48	0.45
Decision Tree				
Accuracy	0.86	0.80	0.80	0.80
Precision	0.61	0.57	0.56	0.56
Recall	0.63	0.59	0.58	0.58
F-1 Score	0.60	0.57	0.55	0.55
Random Forest				
Accuracy	0.87	0.81	0.81	0.82
Precision	0.65	0.60	0.58	0.59
Recall	0.67	0.60	0.61	0.60
F-1 Score	0.62	0.58	0.57	0.57
Gaussian Naïve Bayes				
Accuracy	0.28	0.22	0.46	0.27
Precision	0.27	0.26	0.32	0.29
Recall	0.25	0.25	0.41	0.38
F-1 Score	0.16	0.13	0.22	0.16

Table 1: UNSW-NB15 Performance Comparison (Macro)

IDS-IoT2024 (Macro)				
ML Metric	**Manual**	**DeepSeek V3**	**Gemini 2.5 Pro**	**GPT-5**
K-Nearest Neighbor				
Accuracy	0.88	0.95	0.95	0.95
Precision	0.86	0.88	0.84	0.84
Recall	0.78	0.91	0.92	0.93
F-1 Score	0.79	0.88	0.85	0.85
Decision Tree				
Accuracy	0.81	0.95	0.95	0.95
Precision	0.72	0.89	0.86	0.91
Recall	0.85	0.90	0.91	0.91
F-1 Score	0.73	0.89	0.87	0.91
Random Forest				
Accuracy	0.81	0.95	0.95	0.95
Precision	0.72	0.89	0.86	0.90
Recall	0.85	0.91	0.91	0.91
F-1 Score	0.73	0.90	0.87	0.91
Gaussian Naïve Bayes				
Accuracy	0.30	0.91	0.84	0.87
Precision	0.50	0.85	0.70	0.77
Recall	0.55	0.85	0.75	0.86
F-1 Score	0.33	0.79	0.64	0.74

Table 2: IDS-IoT2024 Performance Comparison (Macro)

Conclusion and Future Work

Generative AI models, though unable to think independently, can draw upon their vast internal knowledge. As a proof of concept, our experiment demonstrates that with specific instructions, GenAI can be "taught" to conduct exploratory analysis and generate data preprocessing code equivalent to human-written code. Future work will focus on: testing new models for ML preprocessing, enhancing instruction language for better teaching, and expanding the types of ML algorithms tested to optimize results.

References

Cote, P. O.; Nikanjam, A.; Ahmed, N.; Humeniuk, D.; and Khomh, F. 2024. Data Cleaning and Machine Learning: A Systematic Literature Review. *Automated Software Engineering*, 31(54).

Dhawas, P.; Dhore, A.; Bhagat, D.; Dorlikar, R. D.; Kukade, A.; and Kalbande, K. 2024. Big Data Preprocessing, Techniques, Integration, Transformation, Normalisation, Cleaning, Discretization, and Binning. In Darwish, D., ed., *Big Data Analytics Techniques for Market Intelligence*, 159–182. IGI Global.

Koppula, M.; and Leo Joseph, L. M. I. 2025. A Real-World Dataset 'IDSIoT2024' for Machine Learning/Deep Learning Based Cyber Attack Detection System for IoT Architecture. In *2025 3rd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT)*, 1757–1764. IEEE.

Koppula, M.; and L.M.I, L. J. 2024. A real time dataset 'IDSIoT2024.'. Accessed: 2025-09-14.

Moustafa, N.; and Slay, J. 2015a. UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set). In *2015 Military Communications and Information Systems Conference (MilCIS)*.

Moustafa, N.; and Slay, J. 2015b. The UNSW-NB15 Dataset. https://research.unsw.edu.au/projects/unsw-nb15-dataset. Accessed: 2025-09-14.