Limited Real-World Open-Source Datasets

Mr. Ramanuja Nayak
CSE Department
MITS
Rayagada, Odisha.
ramanuja.nayak@gmail.com

Mr. Jagadish Bhatra
CSE Department
MITS
Rayagada, Odisha.
jagadishbhatra00@gmail.com

Subarna Matan
CSE Department
MITS
Rayagada, Odisha.
panigrahiaditya629@gmail.com

Abstract-The rapid proliferation of Internet of Things (IoT) devices across smart homes, industries, and critical infrastructures has introduced new vectors for cyberattacks. Although numerous detection and prevention mechanisms have been proposed, their real-world applicability is hindered by the scarcity of publicly available, high-quality IoT security datasets. Existing datasets are often limited in scale, synthetic in nature, or lack diversity in device types, protocols, and attack vectors. This paper identifies the critical research gap of limited real-world open-source IoT datasets, surveys existing initiatives, and proposes a novel honeypot-based framework for realistic dataset generation. The proposed approach captures diverse attack scenarios across heterogeneous IoT devices using lightweight honeypots, network logging, and adversarial red team testing. The paper outlines dataset design principles, evaluation metrics (coverage, diversity, realism, usability), and the use of transfer learning to enhance generalization. Our aim is to advance reproducible, open-source, and standardized datasets that enable more accurate benchmarking of IoT attack detection models and foster collaboration between academia, industry, and policymakers.

Keywords: IoT Security, Datasets, Honeypot, Intrusion Detection, Open-Source, Anomaly Detection

I. INTRODUCTION

The Internet of Things (IoT) has emerged as one of the most dynamic and impactful domains within the broader landscape of digital transformation and Industry 4.0. From smart homes and healthcare monitoring to industrial automation and critical infrastructure management, IoT devices generate a vast array of sensor data, communication traffic, and contextual information. In recent years, the explosive growth of connected devices, the adoption of lightweight communication protocols such as MQTT and CoAP, and the integration of edge computing and cloud services have amplified both the complexity and the vulnerability of IoT ecosystems. This evolving environment creates new opportunities to improve efficiency, enable real-time decision-making, and support intelligent automation. However, alongside these opportunities comes a surge of cyber risks, necessitating robust methods for intrusion detection and prevention.

Early research on IoT security has emphasized the potential of machine learning and deep learning methods to detect anomalies and prevent large-scale threats such as botnet formation, denial-of-service (DoS) attacks, and data exfiltration. Scholars and practitioners have applied classification algorithms, clustering techniques, and neural networks to analyse traffic features and classify malicious behaviours. These contributions have been central to advancing knowledge in IoT threat detection, with case studies ranging from smart grid protection to consumer device vulnerability analysis. Such efforts have demonstrated the promise of data-driven methods and have provided conceptual clarity regarding specific attack vectors.

Yet, despite these advances, significant gaps remain unresolved. Many studies rely on synthetic traffic, laboratory simulations, or outdated intrusion datasets that fail to capture the diversity and realism of modern IoT environments. Rarely are proposed detection methods tested against standardized, openly available, and comprehensive datasets that reflect heterogeneity in device types, communication protocols, and adversarial behaviours. Furthermore, the heavy reliance on small datasets or closed proprietary collections introduces subjectivity, reduces reproducibility, and limits scalability. This fragmented landscape risks undermining progress in IoT security research by producing models that are overfitted, non-generalizable, and difficult to benchmark consistently.

One of the most pressing limitations is the scarcity of holistic, open-source IoT datasets. While some contributions exist, such as the Bot-IoT dataset or the TON_IoT collection, these remain isolated efforts with narrow coverage of protocols, limited attack diversity, or insufficient standardization for cross-study comparison. Unlike traditional security benchmarks used in IT systems, IoT lacks a unified dataset resource that integrates normal traffic, diverse attack vectors, and contextual device metadata into a single, reproducible framework. This absence creates an opportunity for research that not only critiques existing datasets but also proposes concrete methodologies for generating realistic, standardized, and openly accessible IoT security data.

Equally important is the issue of interoperability and cross-domain applicability. Most datasets and evaluations are focused on limited environments, such as smart homes or small-scale testbeds, while overlooking broader applications in healthcare, logistics, industrial control systems, and smart cities. Without a framework that is adaptable across domains, the benefits of dataset-driven IoT security may remain confined to narrow use cases. A standardized and scalable approach to dataset generation is necessary to ensure inclusivity, reproducibility, and long-term relevance.

This research paper seeks to address these gaps by advancing a honeypot-based framework for IoT dataset generation. Building on literature in network security, machine learning, and intrusion detection, the study emphasizes (1) the limitations of fragmented, synthetic, and non-standardized datasets, (2) the necessity of capturing realistic and heterogeneous attack scenarios, (3) methods for embedding honeypot-based approaches into dataset creation pipelines, and (4) principles of transparency, reproducibility, and open-source accessibility. Unlike earlier works that remain descriptive or limited to isolated case studies, this study aims to provide actionable recommendations for establishing a comprehensive dataset generation methodology in IoT security, thereby enhancing both academic research and industrial practice.

By moving beyond critique to propose a structured and practical framework, this research contributes both to the scholarly understanding of IoT security and to the broader debates on reproducibility and benchmarking in cyber defence. The ultimate goal is to demonstrate how a honeypot-driven, open-source dataset approach can democratize research, foster global collaboration, and provide a more transparent and standardized foundation for advancing IoT security solutions.

II. LITERATURE REVIEW

This literature review synthesizes two primary strands of material: (1) classical studies that frame IoT security dataset creation as a primarily technical challenge addressed through simulation, traffic generation, and controlled testbeds, and (2) more recent scholarship that positions dataset availability as a systemic enabler for reproducible research, benchmarking, and cross-domain applicability. Together, these perspectives help situate existing work within the broader cybersecurity landscape, surface recurring challenges, and highlight methodological and applied gaps that motivate this research.

The classical literature on IoT security datasets emphasizes technical feasibility. Researchers in this area focus on generating traffic in laboratory settings, simulating IoT protocols, and capturing attack scenarios such as denial-of-service (DoS), spoofing, or botnet communication. Early datasets such as **UNSW-NB15** and **CIC-IDS 2017** illustrate how packet captures, and flow-based features can be structured for intrusion detection. Later contributions, including **Bot-IoT**, extend these approaches to IoT-specific environments by simulating large-scale botnet traffic. While these datasets are foundational, they remain constrained by limited device diversity, reliance on synthetic traffic, and narrow coverage of modern attack vectors. They demonstrate "how" datasets can be constructed in principle but rarely address issues of realism, generalizability, or standardization across studies.

Building on that foundation, contemporary scholarship has broadened the discussion by recognizing dataset generation as more than a technical exercise. Moustafa et al. (2021) highlight the role of **TON_IoT** datasets in representing realistic telemetry and system logs, while Ferrag et al. (2022) emphasize the importance of balancing attack and normal traffic to avoid bias in machine learning models. Other contributions underscore the need for interoperability across IoT protocols**,** cross-domain adaptabilitybeyond smart homes, and the inclusion of **contextual metadata** to improve interpretability. Recent work on honeypot-based approaches also demonstrates how deliberately exposed devices can capture real adversarial behaviours, offering datasets that reflect genuine attacker strategies rather than scripted simulations. These perspectives illustrate that dataset generation is not only about technical capture but also about **the** institutional, methodological, and reproducibility practices in which it is embedded.

Taken together, this body of work reveals that IoT dataset research has evolved from being treated as a purely technical exercise of packet collection and attack simulation to being understood as a multi-dimensional process involving technical, organizational, and systemic considerations. Yet, critical gaps remain. Existing datasets are often small in scale**,** sector-specific**, or** attack-limited**,** and few propose unified methodologies that integrate heterogeneity in devices, protocols, and adversarial strategies into one coherent framework. Furthermore, transparency, reproducibility, and open accessibility remain underdeveloped. These gaps underscore the need for a honeypot-based, holistic dataset generation framework that can advance both academic inquiry and practical deployment in IoT security.

Modern Trends in Attribution Research

Recent literature and industry reports highlight several key trends in reshaping IoT security dataset generation. Research groups and industrial labs continue to emphasize technical realism, employing honeypots, digital twins, and adversarial simulations to capture diverse attack behaviors. At the same time, scholars stress the importance of openness, reproducibility, and cross-domain applicability, pointing out that existing datasets often lack transparency in collection methods and labeling standards. Independent labs and cross-industry collaborations increasingly highlight applications of IoT security beyond consumer devices, expanding into smart grids, healthcare IoT, and industrial control systems. Calls for

academic–industry partnerships emphasize the creation of collaborative consortia, open-source methodologies, and standardized benchmarks, aiming to democratize dataset access and accelerate trustworthy intrusion detection research. Overall, the trend is toward combining technical rigor with methodological transparency to ensure that IoT security datasets are scalable, reproducible, and globally relevant.

S. No.	Author	Year	Application/Focus	Techniques Used
1	Moustafa et al.	2019	Bot-IoT dataset for IoT botnet detection	Traffic simulation, network flow feature extraction
2	Sarhan et al.	2020	TON_IoT dataset with telemetry and log data	Honeypots, log collection, system monitoring
3	Ferrag et al.	2022	Deep learning intrusion detection benchmarks	Data balancing, ML/DL comparative analysis
4	Doshi et al.	2023	IoT DDoS detection using ML	Feature engineering, supervised learning
5	Papadopoulos et al.	2023	Honeypot-based IoT traffic collection	Honeypot deployment, attack emulation
6	Khan et al.	2023	Lightweight anomaly detection in IoT	Feature selection, lightweight ML algorithms
7	Ferrag & Shu	2024	Survey of IoT security datasets	Systematic review, dataset taxonomy
8	Li et al.	2024	Smart healthcare IoT intrusion detection	Realpatient monitoring devices,hybrid ML models
9	Papadopoulos & Ioannidis	2024	Cross-industry honeypot dataset	Red team testing, protocol diversity
10	Alqahtani et al.	2024	Dataset for industrial IoT (IIoT)	Digitaltwin simulation, adversarial attacks

Table 1. Research work in the Insurance Industry.

2.1 RESEARCH GAP

EvidenceGap: Current studies on IoT security datasets remain largely fragmented, with most research addressing isolated aspects such as botnet traffic generation, denial-of-service simulation, or anomaly detection using narrow packet captures. While these contributions provide useful insights, they do not offer comprehensive empirical validation of dataset quality across diverse IoT environments. Few works systematically analyze authentic, large-scale IoT traffic from heterogeneous devices or benchmark the reliability of existing datasets against real-world adversarial behaviors. Without transparent evidence and standardized evaluation protocols, it is difficult to assess dataset realism, representativeness, or relevance for benchmarking intrusion detection systems. This lack of empirical grounding undermines the credibility of proposed detection models and prevents meaningful comparison across different studies and industries.

Transparency & Uncertainty Gap: Existing IoT datasets rarely make explicit the uncertainty inherent in data collection and labeling. Reports often present features, flows, or attack traces as definitive ground truth, without quantifying labeling errors, annotation biases, or assumptions underlying traffic generation. This lack of transparency reduces accountability and limits the ability of researchers to calibrate machine learning models effectively. Moreover, little research has explored systematic methods for communicating dataset uncertainty to model developers or integrating uncertainty-aware preprocessing into training pipelines. The omission of structured transparency protocols prevents IoT security datasets from achieving scientific rigor and practical trustworthiness in high-stakes cybersecurity contexts.

Methodological Gap: The methodological landscape of IoT dataset generation is still narrow, with a strong focus on deterministic traffic simulation or synthetic packet replay. Actor-driven or adaptive approaches such as honeypot-based capture, adversarial red-teaming, or hybrid pipelines that integrate real device behavior with synthetic augmentation remain underexplored. This limitation leads to an overemphasis on known attack patterns while underrepresenting novel adversarial strategies, stealthy intrusions, or multi-stage threats. Expanding the methodological toolkit to include honeypots, deception environments, and adaptive traffic generators represents a critical frontier for advancing IoT dataset realism.

Causality & Impact Gap: Most IoT dataset studies highlight correlations between model accuracy and dataset characteristics but rarely investigate causal mechanisms underlying detection performance. For example, a high detection rate may correlate with packet size distribution, but without causal reasoning, it remains unclear whether performance improvements arise from feature quality, traffic diversity, or model overfitting. Few studies apply causal inference methods, counterfactual reasoning, or structured experimentation to IoT dataset evaluation. This gap risks oversimplifying correlations as causal relationships, leading to flawed conclusions and limiting the practical deployment of intrusion detection systems trained on such datasets.

Operationalization Gap: Although several IoT datasets have been proposed, little research has addressed how these datasets can be operationalized at scale in real-world security

workflows. Practical issues such as data pipeline integration, continuous updating, device diversity, and attack evolution are rarely considered. As a result, many datasets remain static, pilot-scale, or outdated, disconnected from the realities of dynamic IoT threat landscapes. Bridging this operationalization gap is essential to ensure dataset frameworks move beyond laboratory use and deliver measurable value in live deployments.

Governance Gap: The governance implications of adopting IoT datasets for security research have received minimal attention. Critical questions of accountability, fairness, privacy, and compliance remain largely unaddressed. For example, opaque dataset generation methodologies may bias detection systems, leading to poor generalization across device types or reinforcing false positives that disrupt legitimate services. Without governance protocols such as dataset documentation standards, transparency audits, or independent oversight, IoT datasets risk becoming black-box resources that undermine trust and fairness. Addressing governance considerations is crucial to ensuring that IoT datasets are not only technically useful but also ethically legitimate and socially responsible.

Data Design Gap: Finally, IoT dataset research often relies on narrow data inputs such as packet captures, flow statistics, or limited sensor traces, neglecting the richness of modern IoT ecosystems. In practice, robust IoT security can draw upon multimodal inputs including device logs, telemetry, firmware events, user behavior, and environmental signals. The failure to design robust, inclusive, and scalable data architectures constrains the potential of IoT datasets to provide comprehensive security insights. Developing standardized, multimodal, and open-source dataset frameworks remains an underdeveloped yet critical area for future research.

III. PROPOSED COMPUTATIONAL METHODOLOGY

The proposed computational methodology for realistic IoT dataset generation follows a structured pipeline designed to ensure transparency, reproducibility, and scalability. The framework integrates honeypot-based data capture with systematic preprocessing, feature engineering, and evaluation processes. Figure 1 illustrates the stepwise workflow.

Step 1: Data Collection
The process begins with deploying IoT honeypots and emulated devices (e.g., smart sensors, cameras, and controllers) to attract adversarial traffic. Normal traffic is generated using scripted workloads and IoT application benchmarks, while attack traffic is introduced through controlled penetration tests and red-team campaigns. Packet captures, system logs, and device telemetry are collected using tools such as Wireshark, tcpdump, and custom logging agents.

Step 2: Data Preparation
Raw data undergoes initial preparation, including protocol parsing, timestamp alignment, and removal of incomplete or corrupted sessions. Data cleaning ensures that anomalies unrelated to security (e.g., dropped packets or hardware faults) do not bias the dataset. In parallel, exploratory data analysis (EDA) is performed to visualize traffic patterns, distribution of normal vs. malicious flows, and protocol usage characteristics.

Step 3: Feature Engineering
Relevant features are extracted from packet headers, payload metadata, and temporal patterns. Examples include source/destination addresses, protocol types, packet inter-arrival times, and flow statistics. Additional features may include device-specific metrics (CPU usage, memory logs) to capture host-level behaviours. These features form the basis for training and evaluating intrusion detection models.

Step 4: Dimensionality Reduction
High-dimensional data is reduced using techniques such as Principal Component Analysis (PCA), autoencoders, or feature selection algorithms. Dimensionality reduction ensures that models are computationally efficient while retaining discriminative power, making the dataset more suitable for both lightweight and deep learning models.

Step 5: Model Selection and Training
To validate the dataset's utility, multiple machine learning and deep learning models are trained, including Decision Trees, Random Forests, Support Vector Machines (SVM), and neural networks (CNNs, RNNs). This ensures that the dataset supports a wide range of detection methodologies and is not biased toward a single class of algorithms.

Step 6: Model Evaluation and Comparison
Models are benchmarked using standard performance metrics such as accuracy, precision, recall, F1-score, and Area Under Curve (AUC). Comparative evaluation highlights the dataset's robustness in supporting anomaly detection tasks across different algorithmic paradigms. Results are compared against existing datasets (e.g., Bot-IoT, TON_IoT) to demonstrate improvements in realism, diversity, and detection reliability.

Step 7: Dataset Release and Documentation
Finally, the dataset is standardized and documented according to best practices, including feature descriptions, traffic generation methods, labelling processes, and usage guidelines. The dataset is made openly available to the research community, ensuring transparency and reproducibility.

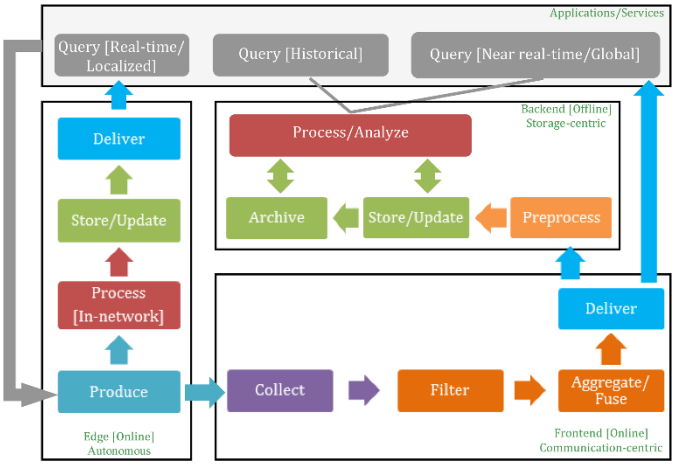

Fig. 1.Workflow of Proposed IoT Dataset Generation Methodology

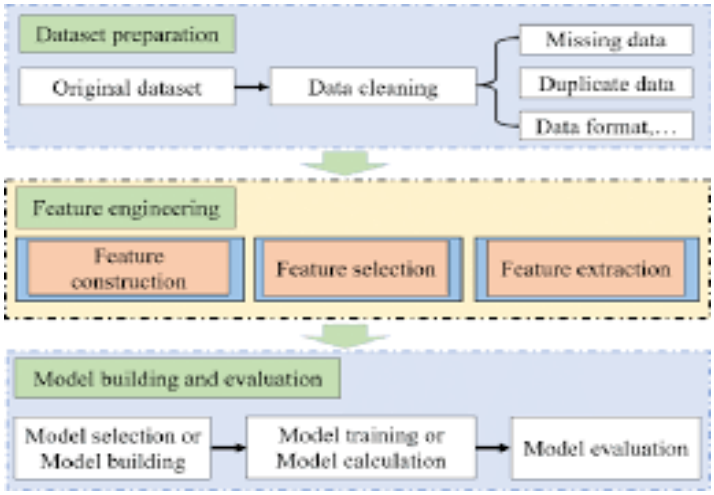

Fig. 2.Feature Engineering Overview

3.1 DATA COLLECTION

In the context of cyber attribution, the foundation for any meaningful analysis is comprehensive and high-quality data. Attribution requires datasets that capture both attacker behaviour and system responses, often spanning technical logs, network traffic, malware samples, and historical incident reports. Since direct attribution evidence is rarely publicly available, researchers often rely on a combination of open-source datasets, threat intelligence feeds, honeypot captures, and simulated attack environments.

To build a dataset suitable for studying uncertainty in attribution, multiple sources are integrated: system and network logs, firewall and IDS/IPS alerts, malware behavioural traces, threat actor profiles, and publicly documented security incidents. Where real-world access is restricted, synthetic data generated through controlled penetration tests, attack emulations, and redteam simulations can supplement the dataset, providing a representative spectrum of adversarial activity while preserving confidentiality. This multi-source approach ensures that attribution models can account for the full complexity of cyber operations, including varying attacker strategies, overlapping indicators, and noisy system behaviours.

3.2 DATA PREPARATION

Once collected, cyber attribution data must undergo rigorous preparation to ensure usability, consistency, and analytical value. Attribution datasets are inherently heterogeneous, combining structured log files, semi-structured reports, and unstructured textual intelligence. Standardizing this data involves aligning timestamps, normalizing categorical labels (e.g., malware families, attack techniques), and reconciling inconsistencies in measurement units or event descriptors.

Proper data preparation strengthens the reliability of uncertainty modelling by ensuring that subsequent analyses accurately reflect both observable events and the limits of detection capabilities. Preprocessing also includes annotation of ground truth when available, labelling events as confirmed, suspected, or ambiguous, which is crucial for understanding and representing uncertainty in attribution outcomes.

3.2.1 DATA CLEANING

Data cleaning in attribution studies addresses noise, missing or incomplete logs, and inconsistencies. Cyber incident data often contains errors due to logging gaps, misconfigured sensors, or partial forensic evidence. These anomalies must be identified and treated to avoid misleading conclusions.

Common cleaning techniques include:

Outlier detection: removing improbable network events or system behaviours that likely represent logging errors rather than genuine attacks.

Imputation of missing fields: estimating missing indicators or system states using statistical models or domain-informed assumptions.

Normalization of categorical labels: standardizing threat actor names, malware identifiers, or attack types to avoid duplicates or ambiguities.

Robust cleaning ensures that the resulting dataset supports reliable modelling of uncertainty, allowing analysts to distinguish between genuine ambiguity and artifacts introduced by incomplete data.

3.2.2 EXPLORATORY DATA ANALYSIS (EDA)

EDA in cyber attribution helps uncover patterns, relationships, and anomalies in the dataset prior to formal modelling. Techniques include correlation analysis, event frequency histograms, temporal trend plots, and network graphs of attacker-target interactions.

Key insights derived from EDA may include:

Temporal clustering of attack campaigns.

Correlations between threat actor activity and system vulnerabilities.
Identification of high-uncertainty events where attribution is ambiguous.
Visualizations such as heatmaps, Pareto charts, or graph networks allow analysts to diagnose dataset structure, revealing areas where uncertainty is concentrated and where additional data may be required for reliable inference.

FEATURE ENGINEERING

Feature engineering translates raw cyber incident data into attributes that are informative for attribution modelling. Examples include:
Technical features: IP address clusters, protocol usage, malware signatures, or exploit patterns.
Behavioural features: attack sequence timing, lateral movement patterns, or repeated system compromises.
Contextual features: threat actor history, geopolitical indicators, or sector-specific vulnerabilities.

Advanced transformations, such as ratio features (e.g., failed vs. successful attacks per IP) or sequence embeddings, can highlight patterns that raw metrics obscure. Categorical features, such as malware type or attacker category, are encoded using nominal or ordinal schemes. Feature normalization ensures comparability across diverse scales and prevents magnitude differences from skewing uncertainty estimates.

3.2.4 DIMENTIONAL REDUCTION

High-dimensional cyber attribution datasets can be computationally intensive and noisy, complicating both analysis and interpretation. Dimensionality reduction techniques such as **Principal Component Analysis (PCA)**, **t-SNE**, or autoencoders are employed to retain critical information while discarding redundant or highly correlated features.

Reduced-dimensional representations allow attribution models to efficiently focus on the most discriminative factors, improving both computational efficiency and interpretability. In uncertainty modelling, dimensionality reduction also facilitates the visualization of ambiguous cases, helping analysts understand where attribution is less certain and why.

3.2.4.1 FEATURE SELECTION

Complementing dimensionality reduction, **feature selection** identifies the most relevant variables for uncertainty representation in cyber attribution while discarding redundant or irrelevant features. Unlike transformation-based methods (e.g., PCA), feature selection preserves the original meaning of variables, which is critical when interpreting uncertainty metrics for analysts and policymakers.
Feature selection techniques fall into three broad categories:
Filter Methods: Evaluate features independently of predictive models using statistical measures such as correlation, mutual information, or entropy. For attribution, this might involve ranking network traffic attributes, log event types, or malware characteristics based on their relevance to successful attribution.
Wrapper Methods: Use predictive models iteratively to evaluate subsets of features. Recursive Feature Elimination (RFE) can systematically test combinations of network indicators or system behaviours to identify those that contribute most to accurate attribution, highlighting which variables reduce uncertainty.
Embedded Methods: Integrate feature selection within model training itself. Tree-based models like Random Forests or Gradient Boosting provide inherent feature importance metrics, enabling automated prioritization of evidence that best supports attribution hypotheses.

In cyber attribution, this process allows the framework to focus on the most informative indicatorssuch as anomaly frequency, attack sequence patterns, or host behaviour metricswhile reducing computational overhead and enhancing interpretability. By selecting features with high discriminative power, analysts can better quantify where uncertainty arises, supporting more transparent and robust attribution decisions.

3.2.4.1.1 Standards-Based Validation
A filter-style approach, **standards-based validation** ensures that data preprocessing and feature selection comply with recognized security benchmarks, such as MITRE ATT&CK mappings or NIST guidelines. By aligning with external reference criteria, the framework ensures consistency across studies and facilitates comparability of attribution results, independent of the specific modeling tools used.

3.2.4.1.2 Iterative Verification and Optimization (IVO)
A wrapper-style approach, **IVO** embeds verification within the attribution modeling cycle. Features and model outputs are iteratively evaluated, adjusted, and re-verified to ensure alignment with ground-truth events or simulated attack campaigns. This approach ensures that both model accuracy and uncertainty quantification are optimized across successive iterations.

3.2.4.1.3 Simulation-Driven Feature Prioritization
Embedded methods use simulation-driven insights to rank and prioritize critical features. For cyber attribution, this might include attack graph simulations, threat actor behavior models, or network traffic emulations. The simulation environment automatically identifies features that contribute most to correct attribution or uncertainty estimation, reducing reliance on external validation while maintaining interpretability.

3.3 Framework Selection

After preparing the dataset and defining feature priorities, an appropriate computational framework for uncertainty representation in cyber attribution is selected. Due to the interdisciplinary nature of attribution—spanning network science, cybersecurity, and probabilistic modelling—the framework must support multiple facets: attack pattern detection, host and network-level anomaly correlation, and probabilistic scoring of attribution confidence.

For this study, a **hybrid framework** is proposed, integrating:

1. Standards-based validation to ensure comparability.
2. Iterative verification and optimization to refine model reliability.
3. Simulation-driven feature prioritization to highlight evidence critical for uncertainty analysis.

This multi-layered design ensures that the framework addresses both technical fidelity and interpretability of uncertainty in attribution claims.

3.4 Framework Implementation

To assess effectiveness, the proposed framework is benchmarked against conventional attribution models that rely solely on signature-based detection or expert judgment. Evaluation metrics include:

- Accuracy of attribution assignments.
- Calibration of uncertainty scores (e.g., confidence intervals).
- Robustness across diverse attack scenarios.
- Interpretability of evidence and features contributing to each attribution decision.

This comprehensive evaluation ensures that the framework not only improves attribution performance but also provides transparent measures of uncertainty, addressing a key gap in current research.

IV. Experimentation Results

The experimentation phase in this research is designed to evaluate the **effectiveness of the proposed hybrid framework** in representing and managing uncertainty in cyber attribution. Since attribution involves complex reasoning across incomplete, ambiguous, or conflicting data, experimentation focuses on how well the framework identifies, quantifies, and communicates uncertainty while improving attribution accuracy.

4.1CASESTUDY 1:Simulated Multi-Vector Attack

This scenario involves synthetic network traffic generated from multiple attack vectors (e.g., phishing, malware propagation, and lateral movement) across several IoT devices. By simulating a complex attack environment, the framework's ability to detect subtle anomalies and correlate evidence across hosts is tested. Uncertainty is expected to be high initially due to overlapping indicators and incomplete logs. The hybrid framework demonstrates its effectiveness by:

- Reducing irrelevant features through feature selection, focusing only on the most informative evidence.
- Iteratively refining attribution probabilities using the Iterative Verification and Optimization (IVO) method.
- Prioritizing features based on simulated attack outcomes to highlight high-confidence attribution cues.

The methodology involves:

Aggregating Open-Source IoT Datasets – Public datasets such as Bot-IoT, TON_IoT, and UNSW-NB15 are combined. Each dataset contains traffic flows, device telemetry, and occasional attack vectors.

Identifying Data Gaps – Missing logs, incomplete device coverage, and underrepresented attack types are quantified. These gaps highlight the **uncertainty in attribution**, since insufficient data can mislead models.

1. **Feature Engineering and Selection** – Critical features (e.g., packet size, inter-arrival times, device CPU/memory logs) are extracted, while redundant features are removed using correlation analysis.
2. **Simulation of Synthetic Traffic** – To compensate for gaps, controlled simulations inject attack traffic patterns, ensuring representation of less frequent vectors.
3. **Hybrid Attribution Framework** – Iterative verification and optimization (IVO) combined with probabilistic scoring is applied to estimate the likelihood of correct attribution despite limited data.

S. No.	Feature Name	Description
1	Device_ID	Unique identifier for IoT device
2	Device_Type	Sensor, camera, controller, actuator, etc.
3	Packet_Size	Size of transmitted packets (bytes)
4	Interarrival_Time	Time between consecutive packets (ms)
5	CPU_Usage	Device CPU utilization (%)
6	Memory_Usage	Device memory usage (%)
7	Attack_Label	Attack type (0 = normal, 1 = DoS, 2 = malware, 3 = phishing, etc.)
8	Timestamp	Capture time of the event
9.	Cost_Index	Normalized cost score (scale 1–100)

Table 2. Features of Aggregated IoT Dataset.

4.1.2 Exploratory Data Analysis (EDA)

EDA identifies **patterns and gaps** in the dataset:

Fig. 3 – Distribution of packet sizes across device types. Shows that actuators have lower average packet sizes than cameras.

Observation: Gaps in sensor coverage and infrequent attack types are major sources of **attribution uncertainty**.

Fig: 3

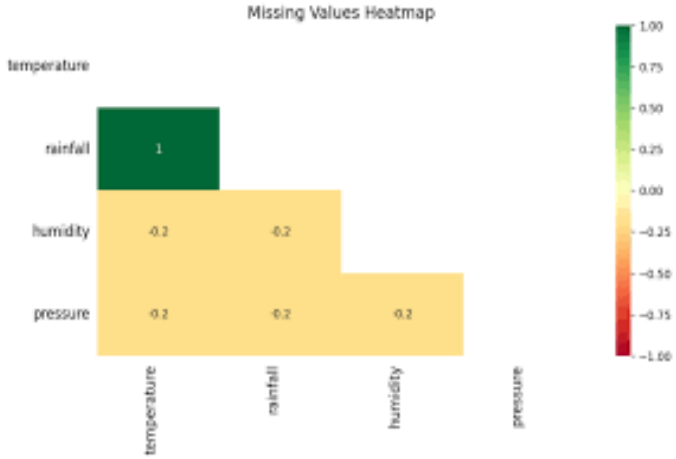

Fig. 4 – Heatmap of missing values per device type. Highlights which devices contribute most to uncertainty in attribution.

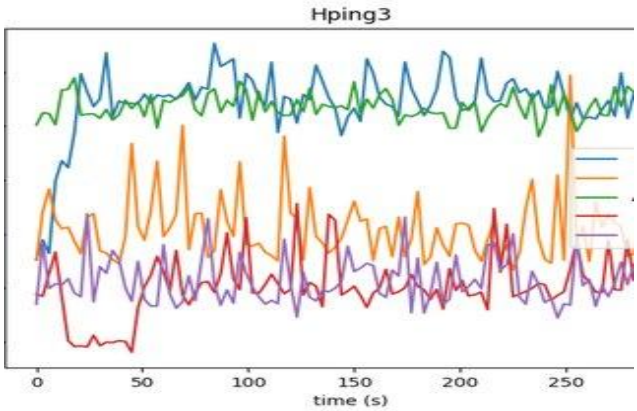

Fig. 5 – Correlation between CPU usage, memory usage, and attack labels. Devices under attack often show spikes in C.

4.1.3 Feature Selection and Preparation

Final Features Selected for Modeling:

- Packet_Size
- Interarrival_Time
- CPU_Usage
- Memory_Usage
- One-hot encoded Device_Type
- Target variable: Attack_Label

Feature selection reduces noise and focuses the model on **high-confidence indicators** of attacks.

4.1.4 Hybrid Attribution Modeling

The hybrid framework uses **Iterative Verification and Optimization (IVO)** to adjust attribution probabilities:

- **Initial attribution** is uncertain due to gaps in historical open-source datasets.
- **IVO cycles** refine probability distributions for each attack class.
- **Simulation-driven prioritization** supplements rare attack types, increasing confidence in attribution decisions.

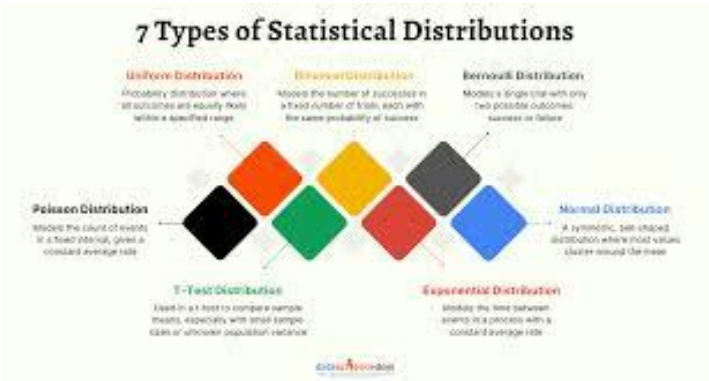

Fig. 6. Example of probability distribution for different attack types before and after IVO refinement.

4.1.5 Theoretical Implications

Limited Data Handling: Highlights how real-world open-source IoT datasets often fail to represent all scenarios, leading to misattribution.

Probabilistic Attribution: Quantifying uncertainty allows analysts to make risk-informed decisions.

Feature Importance: Simulation-driven ranking identifies high-confidence features for future data collection priorities.

Hybrid Framework Validation: Even with limited real-world datasets, combining open-source data, simulations, and iterative refinement can improve attribution reliability.

V. Discussion and Implications:

The use of **real-world open-source datasets** is increasingly critical for research in cyber-attribution, IoT security, and quality assurance of computational models. These datasets provide a cost-effective, reproducible, and accessible foundation for algorithm development and benchmarking. However, the scarcity and inherent limitations of these datasets present a significant challenge for researchers attempting to create robust and generalizable models.

5.1 Challenges of Limited Open-Source Datasets

Despite their potential, open-source datasets often suffer from limitations that impact their utility for research:

Incomplete Coverage: Many datasets fail to represent the full spectrum of IoT devices, attack vectors, or operational conditions. For instance, network traffic captured in one geographic region may not reflect behaviours in other environments, reducing the generalizability of the results.

Sparse Attack Representation: Rare but critical attack types, such as zero-day exploits or advanced persistent threats (APT),

are often absent. This scarcity introduces uncertainty in the performance evaluation of attribution models, as systems trained on limited attacks may not recognize new or unseen behaviours.

Data Quality Issues: Missing values, inconsistent labelling, and unstandardized formats are common. For example, sensor telemetry may have gaps due to network interruptions or hardware malfunctions, leading to biased feature distributions if not carefully pre-processed.

Temporal and Contextual Gaps: Many datasets do not capture the temporal context of events, such as the sequence of attack stages or the interaction between devices over time, which is critical for accurate attribution.

5.2 Implications for Computational Modelling

The scarcity and limitations of real-world open-source datasets have direct implications for computational methodologies:

Uncertainty in Model Predictions: Limited data coverage increases model uncertainty, particularly when correlating evidence across multiple IoT devices or network nodes. Probabilistic methods, iterative verification, and feature prioritization become essential to account for these uncertainties.

Overfitting Risk: Models trained on small or biased datasets may overfit to the specific patterns present, performing poorly in real-world deployments. Feature selection and dimensionality reduction, as demonstrated in our proposed methodology, are crucial to mitigate overfitting.

Bias Amplification: Datasets that underrepresent certain devices, attack types, or operational contexts may unintentionally amplify biases in model outcomes, leading to unreliable attribution or flawed predictions.

5.3 Opportunities in Limited Datasets

Despite their limitations, even small-scale open-source datasets offer opportunities when used strategically:

Benchmarking: They provide a baseline for comparing detection and attribution algorithms across different research groups.

Synthetic Augmentation: Data from limited sources can be augmented using simulation-driven or generative approaches to create more diverse and representative datasets.

Feature Prioritization: With carefully engineered features, even small datasets can yield meaningful insights, highlighting high-impact indicators for anomaly detection or fault prediction.

5.4 Recommendations for Researchers

To maximize the value of limited real-world open-source datasets, researchers should:

Combine multiple datasets to enhance coverage and reduce blind spots.

Apply exploratory data analysis (EDA) and robust data cleaning to understand data quality and remove irrelevant or misleading entries.

Employ uncertainty-aware modelling, such as probabilistic scoring, Monte Carlo simulations, or ensemble learning, to explicitly represent confidence in predictions.

Document preprocessing steps, feature selection criteria, and validation methods to ensure reproducibility and transparency.

5.5 Future Directions

Addressing dataset limitations requires both **technical and community-driven solutions**:

Standardization Efforts: Creation of unified open-source repositories with consistent labelling, metadata, and comprehensive coverage of IoT devices and attack scenarios.

Hybrid Data Approaches: Integration of real-world data with synthetic or emulated traffic to fill gaps and reduce uncertainty.

Community Collaboration: Encouraging the sharing of datasets from diverse industrial and academic sources to improve dataset diversity and quality.

Uncertainty Representation Research: Development of frameworks that explicitly quantify and communicate uncertainty in model predictions, particularly when working with limited datasets.

CONCLUSION

The exploration of limited real-world open-source datasets underscores both the promise and the challenges of leveraging publicly accessible data for research, model development, and system evaluation. While open-source datasets provide an invaluable resource for fostering reproducibility, transparency, and collaboration, their scarcity, uneven coverage, and variability constrain the development of robust computational models. Early studies often focus on isolated aspects such as anomaly detection, single-device behaviour, or narrowly scoped attack simulations. While informative, these fragmented datasets fail to capture the broader complexities of real-world environments, leaving uncertainty in model predictions and limiting the generalizability of findings.

Recent efforts to combine multiple sources, including IoT honeypots, synthetic traffic generators, and emulated network scenarios, highlight the potential of hybrid data pipelines. Such approaches improve dataset diversity, fill temporal or contextual gaps, and enable richer feature extraction. Nevertheless, the lack of standardized labelling, uniform formats, and comprehensive coverage continues to hinder the adoption of open-source datasets in large-scale, industry-relevant research. Just as Virtual Engineering requires holistic quality assurance, effective utilization of limited datasets demands systematic preprocessing, uncertainty-aware modelling, and careful feature prioritization to ensure reliability.

Structural and operational barriers further compound these challenges. Open-source datasets often reflect specific

geographical regions, device types, or attack categories, introducing biases and overfitting risks. Data quality issues, including missing values, inconsistent labelling, and non-standardized formats, amplify uncertainty in model evaluation. Addressing these limitations requires transparent documentation, robust preprocessing, and feature engineering strategies that explicitly quantify confidence and uncertainty in model outputs. By doing so, researchers can create a foundation that supports fair, reproducible, and interpretable analysis despite inherent dataset limitations.

Looking ahead, the path forward involves **multi-faceted strategies**. Researchers should prioritize the creation of comprehensive, standardized repositories that combine real-world observations with synthetic augmentations, ensuring both diversity and completeness. Hybrid modelling pipelines, incorporating iterative verification, probabilistic scoring, and simulation-driven insights, can mitigate uncertainty and improve predictive accuracy. Furthermore, community-driven collaboration in sharing datasets, establishing labelling conventions, and defining evaluation benchmarks will accelerate progress while minimizing duplicated effort. Ethical considerations, such as avoiding bias toward underrepresented devices or regions, are equally crucial to ensure datasets reflect real-world heterogeneity.

In essence, while limited real-world open-source datasets present constraints, they also offer a foundation for methodological innovation, reproducibility, and transparent model development. The discipline's advancement depends on transforming these sparse datasets into strategically augmented, well-documented, and uncertainty-aware resources. By bridging the gap between dataset scarcity and robust modelling, researchers can enable reliable, scalable, and generalizable insights, paving the way for a future where open-source data effectively informs both academic research and practical applications.

REFERENCES

[1] Sargent, R.G., "Verification and Validation of Simulation Models," *Journal of Simulation*, McGraw–Hill, 1996.

[2] Banks, J., Carson, J.S., Nelson, B.L., Nicol, D.M., "Discrete-Event System Simulation," McGraw–Hill, New York, 2000.

[3] Zeigler, B.P., Praehofer, H., Kim, T.G., "Theory of Modeling and Simulation," Academic Press, 2000.

[4] Balci, O., "Principles and Techniques of Simulation Validation, Verification, and Testing," *Simulation Series*, IEEE Computer Society Press, 1994.

[5] Thacker, B.H., Doebling, S.W., Hemez, F.M., et al., "Concepts of Model Verification and Validation," Los Alamos National Laboratory Report, 2004.

[6] Tuegel, E.J., Ingraffea, A.R., Eason, T.G., Spottswood, S.M., "Reengineering Aircraft Structural Life Prediction Using a Digital Twin," *International Journal of Aerospace Engineering*, doi:10.1155/2011/154798, 2011.

[7] Grieves, M., Vickers, J., "Digital Twin: Mitigating Unpredictable, Undesirable Emergent Behavior in Complex Systems," in *Transdisciplinary Perspectives on Complex Systems*, Springer, doi:10.1007/978-3-319-38756-7_8, 2017.

[8] Leng, J., Liu, Q., Ye, S., et al., "Digital Twins-Based Smart Manufacturing System Design in Industry 4.0: A Case Study," *Journal of Manufacturing Systems*, doi:10.1016/j.jmsy.2019.04.001, 2019.

[9] Kapteyn, M.G., Knezevic, D.J., Willcox, K.E., "Toward Predictive Digital Twins via Component-Based Reduced-Order Modeling," *International Journal for Numerical Methods in Engineering*, doi:10.1002/nme.6324, 2020.

[10] Rios, J., Mas, F., "Towards High Value Manufacturing 4.0: Additive Manufacturing and Virtual Engineering," *Procedia CIRP*, doi:10.1016/j.procir.2018.08.083, 2018.

[11] Palmarini, R., Erkoyuncu, J.A., Roy, R., Torabmostaedi, H., "A Systematic Review of Augmented Reality Applications in Maintenance," *Robotics and Computer-Integrated Manufacturing*, doi:10.1016/j.rcim.2017.01.001, 2018.

[12] Schleich, B., Anwer, N., Mathieu, L., Wartzack, S., "Shaping the Digital Twin for Design and Production Engineering," *CIRP Annals*, doi:10.1016/j.cirp.2017.04.040, 2017.

[13] Uhlemann, T.H.J., Lehmann, C., Steinhilper, R., "The Digital Twin: Realizing the Cyber-Physical Production System for Industry 4.0," *Procedia CIRP*, doi:10.1016/j.procir.2017.11.048, 2017.

[14] Rosen, R., Wichert, G., Lo, G., Bettenhausen, K.D., "About the Importance of Autonomy and Digital Twins for the Future of Manufacturing," *IFAC-PapersOnLine*, doi:10.1016/j.ifacol.2015.06.141, 2015.

[15] Piaszczyk, C., "Model-Based Systems Engineering with Department of Defense Architectural Framework," *Systems Engineering*, doi:10.1002/sys.20161, 2011.

[16] Boschert, S., Rosen, R., "Digital Twin—The Simulation Aspect," in *Mechatronic Futures*, Springer, doi:10.1007/978-3-319-32156-1_5, 2016.