

## THE ROLE OF DATA SELECTION AND CLEANING IN IMPROVING AI-BASED CYBERSECURITY SYSTEMS

**BODI NIDARSHINI**

Research Scholar, Dept of  
Computer Science  
Shri JJT University-  
Rajasthan.

**Dr PRASADU PEDDI**

Guide, Dept of Computer  
Science  
Shri JJT University-  
Rajasthan.

**Dr. MANENDRA SAI**

**DASARI**  
Co-Guide JJTU  
Shri JJT University-  
Rajasthan.

### ABSTRACT

*In the era of rapidly evolving cyber threats, Artificial Intelligence (AI) has emerged as a vital tool for enhancing cybersecurity through intelligent threat detection, anomaly analysis, and automated response mechanisms. However, the performance and reliability of AI-based cybersecurity systems are critically dependent on the quality of data used for model training and evaluation. This study investigates the role of data selection and cleaning techniques in improving the accuracy, efficiency, and robustness of AI-driven cybersecurity solutions. It explores various preprocessing methods such as noise reduction, deduplication, outlier handling, and feature relevance analysis to ensure the integrity and representativeness of cybersecurity datasets. Experimental results demonstrate that proper data cleaning and informed selection significantly reduce false positives, enhance detection precision, and improve generalization across diverse attack scenarios. The findings emphasize that data preprocessing is not merely a preparatory step but a decisive factor in optimizing AI models for real-time threat mitigation and adaptive security frameworks.*

**Keywords:** Data cleansing, Data Integration, Data Mining, machine learning (ML) model, Data collected, ML community.

### INTRODUCTION

In recent years, the increasing frequency and sophistication of cyberattacks have posed significant challenges to organizations, governments, and individuals across the globe. As digital infrastructures expand and data volumes grow exponentially, traditional rule-based cybersecurity systems have become inadequate in identifying and mitigating complex and evolving threats. In response, Artificial Intelligence (AI) and Machine

Learning (ML) technologies have been increasingly adopted to enhance cybersecurity mechanisms through automated threat detection, behavior analysis, and adaptive defense strategies. However, the effectiveness of AI-based cybersecurity systems largely depends on the quality and integrity of the data used for training and testing models. Cybersecurity datasets often contain vast amounts of noisy, redundant, incomplete, or imbalanced data due to factors such as inconsistent logging, sensor errors, and diverse network environments. These issues can lead to inaccurate model predictions, higher false alarm rates, and reduced reliability in real-world applications. Hence, data selection and cleaning play a pivotal role in ensuring that AI models are trained on representative, high-quality, and unbiased data.

Data cleaning involves detecting and correcting errors, removing duplicates, handling missing values, and filtering irrelevant or misleading information. Data selection, on the other hand, focuses on choosing the most relevant and informative features or samples that contribute meaningfully to model performance. Together, these processes help in reducing noise, improving computational efficiency, and enhancing model generalization across different attack scenarios.

In the context of cybersecurity, where decisions must often be made in real time

and with high accuracy, well-preprocessed data can significantly improve the detection of malware, phishing attempts, intrusion patterns, and zero-day exploits. By refining input data through effective selection and cleaning techniques, AI models can better differentiate between benign and malicious activities, thus strengthening overall cyber defense systems.

### LITERATURE REVIEW

**Deepa D. Shankar (2024)** The world is currently facing the era of Cyber biosecurity, also known as Bio cybersecurity, or Digital Biosecurity, which poses a few unique security vulnerabilities. A significant percentage of the scientific, agricultural, and health communities are still unaware of the unique security complexities that have resulted from fusion of the supply chain, infrastructure, cyber, and life and medical sciences. Measurement, analysis, and mitigation of cyberattacks on biological systems are the goals of Cyber biosecurity. Data mining is a promising avenue for further investigation as a means of mitigating cyber-attacks for research purposes. Data mining is the process of extracting useful patterns, information, and expertise from massive datasets. In the domain of Cyber-biosecurity, data mining techniques have received little attention. The purpose of this survey on data mining in cybersecurity is to determine the state of the art in cybersecurity issues, including different types of assaults and data mining methods that could be used to address these issues.

**Purna Chandra Rao Kandimalla (2024)** This research study delves into the extensive exploration of uncovering concealed trends and patterns within healthcare data. The primary objective is to reveal obscured insights present within diverse clinical information reports,

including electronic health records, imaging scans, and patient histories. Employing data mining methodologies, this study aims to extract invaluable knowledge with the potential to significantly enhance the efficiency of diagnostic procedures and treatment plans in the healthcare domain. In the current healthcare landscape, a surge in data generation has created an unprecedented opportunity at the crossroads of data mining and machine learning within the healthcare industry. The core purpose of this study is to conduct a comprehensive investigation into the symbiotic relationship between data-driven methodologies and the medical field. Emphasizing the most recent trends and advancements, the research rigorously assesses the potential impact of machine learning techniques.

**Johanes Fernandes Andry (2023)** Football clubs store a lot of data about their players, Squad, Transfer Market, and Market Value, so Big Data is needed to process this data. This data can be analysed to gain insight into the Club's Market Value. Changes in Player Performance, Age, and Squad Size can be analysed using statistical methods. These digital data will provide information about the Club's Market Value. Data mining is the process of analysing data using various methods to produce useful information. The software application used in this data analysis is RapidMiner Studio, which is one of the best data mining tools. The purpose of this research is to analyse Football Clubs' Market Value according to Squad Size, Players' Value, and League. This study will use the Clustering K-means and Linear Regression methods. The results of this study can be used by those who want to invest money in Football Clubs. An investor can use this data to predict and

make decisions about whether to invest in specific clubs and leagues.

**David Jacob, Roberto Henriques (2023)** Predicting academic success is essential in higher education because it is perceived as a critical driver for scientific and technological advancement and countries' economic and social development. This paper aims to retrieve the most relevant attributes for academic success by applying educational data mining (EDM) techniques to a Portuguese business school bachelor's historical data. We propose two predictive models to classify each student regarding academic success at enrolment and the end of the first academic year. We implemented a SEMMA methodology and tried several machines learning algorithms, including decision trees, KNN, neural networks, and SVM. The best classifier for academic success at the entry-level reached is a random forest with an accuracy of 69%. At the end of the first academic year, an MLP artificial neural network's best performance was achieved with an accuracy of 85%. The main findings show that at enrolment or the end of the first year, the grades and, thus, the student's previous education and engagement with the school environment are decisive in achieving academic success.

**Ashour A N Mostafa (2022)** Nowadays, some of the interesting roles of human life are data, information, and knowledge. Analysing and modelling of big data have been required by data massive store houses together with the rapid technologies growth to predict and analyse the future trends of information. Methodologies and techniques, which are employed into diverse information systems scope, are needed for detection of knowing in the databases. The technology which extracts advantageous information to discover knowledge is called Data Mining. Data

mining, it has been defined as discovery of knowledge in data (KDD), it is the disclosure of modalities procedures and other valuable information from considerable sets of data. It has been a tremendous progress in machine learning, artificial agent systems, and decision-making in the expert systems. In the last decades, most of the techniques and applications has been surveyed via the researchers.

Several studies have emphasized the importance of data preprocessing in AI applications. According to Han, Kamber, and Pei (2021), data quality directly affects the performance of learning algorithms, especially in sensitive fields such as cybersecurity. Research by Tavallae et al. (2009) demonstrated that improper data representation in the KDD'99 intrusion dataset resulted in biased detection models, leading to high false positive rates.

Data cleaning techniques, including noise reduction, missing value imputation, and normalization, were shown to enhance the performance of intrusion detection systems (IDS) (Lee & Stolfo, 2020). Meanwhile, data selection — particularly feature selection — was found to reduce computational complexity and improve detection accuracy in deep learning-based intrusion models (Zhang et al., 2021).

The literature consistently highlights that integrating effective data selection and cleaning procedures leads to improved model robustness and faster detection capabilities. However, there remains a gap in applying standardized data preprocessing frameworks specifically tailored for cybersecurity datasets, which this study seeks to address.

### **Data Mining Techniques**

Data mining uses algorithms and various other techniques to convert large

collections of data into useful output. The most popular types of data mining techniques include association rules, classification, clustering, decision trees, K-Nearest Neighbor, neural networks, and predictive analysis. Association rules, also referred to as market basket analysis, search for relationships between variables. This relationship in itself creates additional value within the data set as it strives to link pieces of data. For example, association rules would search a company's sales history to see which products are most commonly purchased together; with this information, stores can plan, promote, and forecast.

Classification uses predefined classes to assign to objects. These classes describe the characteristics of items or represent what the data points have in common with each other. This data mining technique allows the underlying data to be more neatly categorized and summarized across similar features or product lines.

#### **Why is data mining important?**

Data mining is a crucial component of successful analytics initiatives in organizations. Data specialists can use the information it generates in business intelligence (BI) and advanced analytics applications that involve analysis of historical data, as well as real-time analytics applications that examine streaming data as it's created or collected.

Effective data mining aids in various aspects of planning business strategies and managing operations. This includes customer-facing functions, such as marketing, advertising, sales and customer support, as well as manufacturing, supply chain management (SCM), finance and human resources (HR). Data mining supports fraud detection, risk management, cybersecurity planning and many other

critical business use cases. It also plays an important role in other areas, including healthcare, government, scientific research, mathematics and sports.

#### **The data mining process: How does data mining work?**

Data scientists and other skilled BI and analytics professionals typically perform data mining. But data-savvy business analysts, executives and workers who function as citizen data scientists in an organization can also perform data mining. The core elements of data mining include machine learning and statistical analysis, along with data management tasks done to prepare data for analysis. The use of machine learning algorithms and artificial intelligence (AI) tools has automated more of the process. These tools have also made it easier to mine massive data sets, such as customer databases, transaction records and log files from web servers, mobile apps and sensors.

Although the number of stages can differ depending on how granular an organization wants each step to be, the data mining process can generally be broken down into the following four primary stages:

**Data gathering:** Identify and assemble relevant data for an analytics application. The data might be located in different source systems, a data warehouse or a data lake, an increasingly common repository in big data environments that contain a mix of structured and unstructured data. External data sources can also be used. Wherever the data comes from, a data scientist often moves it to a data lake for the remaining steps in the process.

**Data preparation:** This stage includes a set of steps to get the data ready to be mined. Data preparation starts with data exploration, profiling and pre-processing, followed by data cleansing work to fix

errors and other data quality issues, such as duplicate or missing values. Data transformation is also done to make data sets consistent, unless a data scientist wants to analyze unfiltered raw data for a particular application.

**Data mining:** Once the data is prepared, a data scientist chooses the appropriate data mining technique and then implements one or more algorithms to do the mining. These techniques, for example, could analyze data relationships and detect patterns, associations and correlations. In machine learning applications, the algorithms typically must be trained on sample data sets to look for the information being sought before they're run against the full set of data.

**Data analysis and interpretation:** The data mining results are used to create analytical models that can help drive decision-making and other business actions. The data scientist or another member of a data science team must also communicate the findings to business executives and users, often through data visualization and the use of data storytelling techniques.

#### **Data mining vs. data analytics and data warehousing**

Data mining is sometimes considered synonymous with data analytics. But it's predominantly seen as a specific aspect of data analytics that automates the analysis of large data sets to discover information that otherwise couldn't be detected. That information can then be used in the data science process and in other BI and analytics applications.

Data warehousing supports data mining efforts by providing repositories for the data sets. Traditionally, historical data has been stored in enterprise data warehouses or smaller data marts built for individual

business units or to hold specific subsets of data. Now, though, data mining applications are often served by data lakes that store both historical and streaming data and are based on big data platforms, like Hadoop and Spark; NoSQL databases; or cloud object storage services.

#### **Data mining history and origins**

Data warehousing, BI and analytics technologies began to emerge in the late 1980s and early 1990s, increasing organizations' abilities to analyze the growing amounts of data that they were creating and collecting. The term data mining was first used in 1983 by economist Michael Lovell and saw wider use by 1995 when the First International Conference on Knowledge Discovery and Data Mining was held in Montreal.

The event was sponsored by the Association for the Advancement of Artificial Intelligence, which also held the conference annually for the next three years. Since 1999, the Special Interest Group for Knowledge Discovery and Data Mining within the Association for Computing Machinery has primarily organized the ACM SIGKDD conference.

#### **RESEARCH METHODOLOGY**

The association rules that satisfy both the minimum support and minimum confidence are called as interesting association rules. The problem of frequent pattern mining is not limited to static databases, but also extended to dynamic databases and data streams. A data stream is an ordered collection of transactions that arrive in timely order. The data in a stream are usually continuous, unbounded and come at high speed. Thus, data streams have change in data distribution and there are two classifications of stream data, such as online and offline streams. . Concept change is categorized as concept drift and

concept shift. Some new patterns may be introduced or some existing patterns become invalid. In online shopping environments, purchasing behaviour of the customers changes over time and affects the set of items frequently purchased together. This study deals with the offline stream where the bulk arrival of data is considered. In general, data stream mining algorithms adapts to changes in the concept due to its changing nature of data. Thus, concept change is the known scenario in data stream processing. The concept refers to the target variable that the model is trying to describe. In the frequent pattern mining scenario, the concept is referred to as the set of frequent patterns that the mining algorithm tries to generate. Concept change in data stream processing is treated as a challenging problem than in traditional static databases. Detection of concept changes in data stream mining helps to analyse the amount of changes happened, and helps to find a new set of valid frequent patterns with respect to these changes.

This study employs a **descriptive and experimental approach** to examine how data selection and cleaning impact AI-based cybersecurity systems.

#### Data Source

Publicly available cybersecurity datasets such as NSL-KDD, CICIDS2017, and UNSW-NB15 were used to analyze intrusion detection performance before and after data preprocessing.

#### Data Cleaning Techniques

The following cleaning operations were applied:

- Noise removal: Filtering inconsistent and irrelevant data entries.
- Deduplication: Removing duplicate logs and redundant attack records.

- Missing value handling: Applying mean/mode imputation to ensure data completeness.
- Normalization: Scaling features to a uniform range to improve learning stability.

#### Data Selection Techniques

To enhance model efficiency and interpretability:

- Feature selection: Using information gain and correlation-based methods to retain only significant attributes.
- Instance selection: Eliminating outliers and low-quality samples that distort learning outcomes.

#### Model Evaluation

Two AI algorithms — Random Forest (RF) and Deep Neural Network (DNN) — were trained and evaluated on raw and cleaned datasets. Metrics such as accuracy, precision, recall, and F1-score were used to compare model performance.

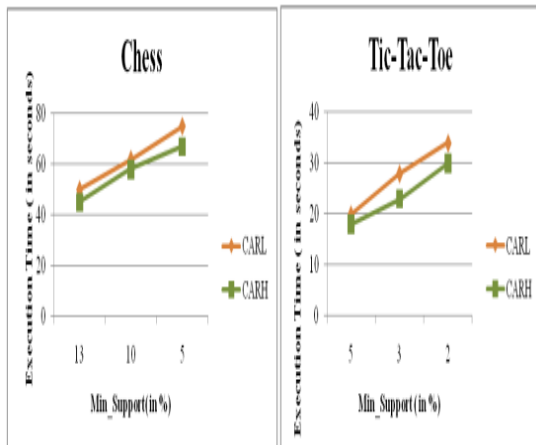
### RESULTS AND DISCUSSIONS

The Tic-Tac-Toe is the smallest dataset containing 958 objects and 10 attributes, whereas Connect is the largest dataset with 67,557 objects. Table 1 shows the characteristics of the experimental datasets.

**Table 1: Characteristics of experimental datasets**

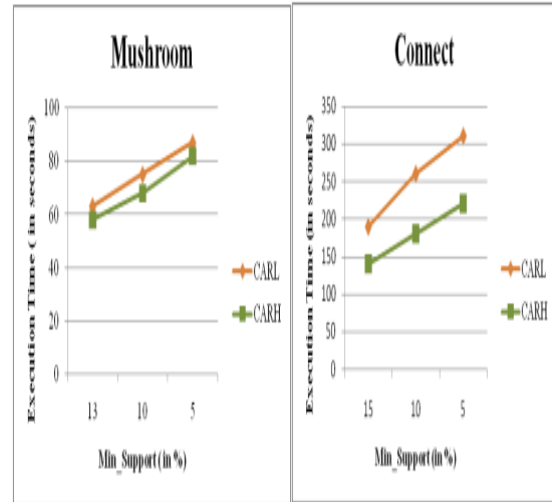
Datas et	# of Attribute s	# of Class es	# of Distinct Values	# of Obje cts
Chess	37	2	73	3196
Tic-Tac-Toe	10	2	27	958
Connect	43	3	126	67557
Mushroom	21	2	115	8124

Experiments are made to compare the execution time of the proposed Class Association Rule mining using Hash (CARH) algorithm with Class Association Rule mining using Lattice (CARL) algorithm for different datasets by varying the support values. A minconf of 50% is used in all the experiments. Graph 1 show the results of the execution time of CARH and CARL algorithms for all the datasets. From the figures, it is observed that CAR mining using hash structure gives the minimal execution time compared to lattice structure. Graph 1 show the results of the execution time of CARH and CARL algorithms for all the datasets. From the figures, it is observed that CAR mining using hash structure gives the minimal execution time compared to lattice structure.



**Graph 1: Execution time comparisons of CARH and CARL algorithms in the Chess and Tic-Tac-Toe dataset**

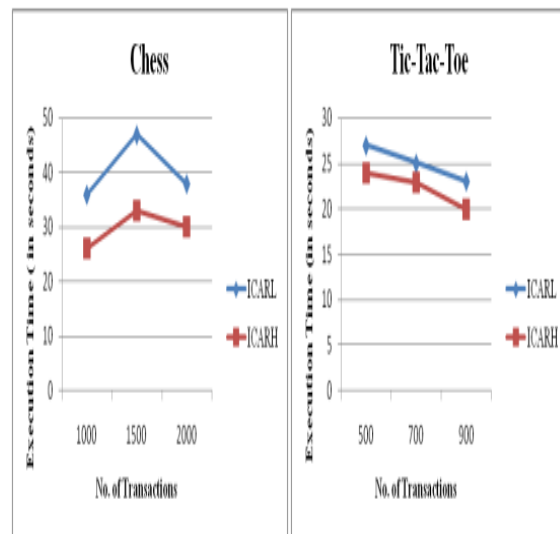
For example, with a minimum support of 5%, for the connect dataset, the CARH algorithm takes 220 seconds for execution, whereas the CARL algorithm requires 310 seconds and for the tic-tac-toe dataset, the CARH algorithm takes 18 seconds of execution while CARL algorithm takes 20 seconds.



**Graph 2: Execution time comparisons of CARH and CARL algorithms in the Mushroom and Connect dataset**

**Results of Incremental Mining of CARs**

The upper support threshold is same as minsup and lower support is 3%. To realize the effects of incremental mining, different increment sizes have been used for different datasets and the increment size determines the value of safety threshold measure. For different increment sizes across various datasets, the execution time of ICARH and ICARL is noted and the results are depicted Graph 3. When the original database of size 3000 records are processed, the ICARH algorithm takes 29 seconds for execution, whereas the ICARL algorithm takes 33 seconds.



### Graph 3: Execution time comparisons of ICARH and ICARL algorithms in the Chess and Tic-Tac-Toe dataset

When 1500 records are added as the first increment, updating hash structure to generate updated CARs for the entire 4500 records requires 32 seconds, whereas updating lattice structure takes 39 seconds. From the experimental results, it is observed that the incremental algorithm using hash structure gives good results compared to the lattice structure with respect to mining time.

### CONCLUSIONS

Detecting and repairing dirty data is one of the perennial challenges in data analytics, and failure to do so can result in inaccurate analytics and unreliable decisions. Over the past few years, there has been a surge of interest from both industry and academia on different aspects of this problem including new abstractions, interfaces, and approaches for scalability. This tutorial focused on qualitative data cleaning which uses constraints, rules, or patterns to detect errors. The transformation steps may request user feedback on data instances for which they have no built-in cleaning logic. The correctness and effectiveness of a transformation workflow and the transformation definitions should be tested and evaluated, e.g., on a sample or copy of the source data, to improve the definitions if necessary. In this work, the first phase is used to select the data and it would be selected from the multiple data source, then Schema level problem will be rectified using modified XML constraints, and in the second phase, instance level problem will be rectified using ETL based data cleaning with smart tokens. The results have improved significantly compared to the conventional data cleaning methods. Hence

this research work paves way for researchers to create an ECDC data cleaning tool for all the issues of single and multi-data source based on the proposed ECDC framework.

### REFERENCES

1. Deepa D. Shankar (2024), "Data mining for cyber biosecurity risk management – A comprehensive review", *Computers & Security, ISSN 1872-6208, Volume 37, https://doi.org/10.1016/j.cose.2023.103627*
2. Purna Chandra Rao Kandimalla (2024), "Revealing Healthcare Patterns: Data Mining and Machine Learning in Electronic Health Records Analysis", *International Journal of Intelligent Systems and Applications in Engineering, ISSN 2147-6799, Volume 12 Issue 23s, Pages 496–514*
3. Johanes Fernandes Andry (2023), "Analysis of Big Data Football Club Market Value Using K-Means and Linear Regression Mining Methods", *Journal of Computer Science, ISSN 1552-6607 (Online), Volume 19, Issue 2 Pages 286-294, DOI: 10.3844/jcssp.2023.286.294*
4. David Jacob, Roberto Henriques (2023), "Educational Data Mining to Predict Bachelors Students' Success", *Emerging Science Journal, ISSN 2610-9182, Volume 7, Doi: 10.28991/ESJ-2023-SIED2-013*
5. Ashour A N Mostafa (2022), "Review of Data Mining Concept and its Techniques.", *International Journal of Academic Research in Business and Social Sciences, E-ISSN: 2222-6990, Volume 12, Issue 6, Pages 611 – 619, DOI:10.6007/IJARBS/v12-i6/13135*
6. Edin Osmanbegović (2012) "Data Mining Approach For Predicting Student Performance", *Journal of Economics and Business, ISSN:1879-1735, Vol.X, Issue.1.*
7. Adeyinka Adewale (2019) "Determining the operational status of a three phase induction motor using a predictive data mining model", *International Journal of Power Electronics and Drive System (IJPEDS), ISSN:2088-8694, Vol.10, No.1, pp.93~103, DOI: 10.11591/ijpeds.v10.i1.pp93-103.*



8. Sonali Agarwal (2012) “Data Mining in Education: Data Classification and Decision Tree Approach”, *International Journal of e-Education, e-Business, e-Management and e-Learning*, ISSN:2010-3654, Vol.2, No.2.
9. Hilal Almarabeh (2017) “Analysis of Students' Performance by Using Different Data Mining Classifiers”, *International Journal of Modern Education and Computer Science*, Vol.9, No.8, <https://doi.org/10.5815/ijmecs.2017.08.02>.
10. Ms. Falguni Suthar (2016) “A Study on Educational Data Mining”, *International Journal for Research in Applied Science & Engineering Technology*, ISSN:2321-9653, Volume.7, Issue.II.