Conference Presentation

By

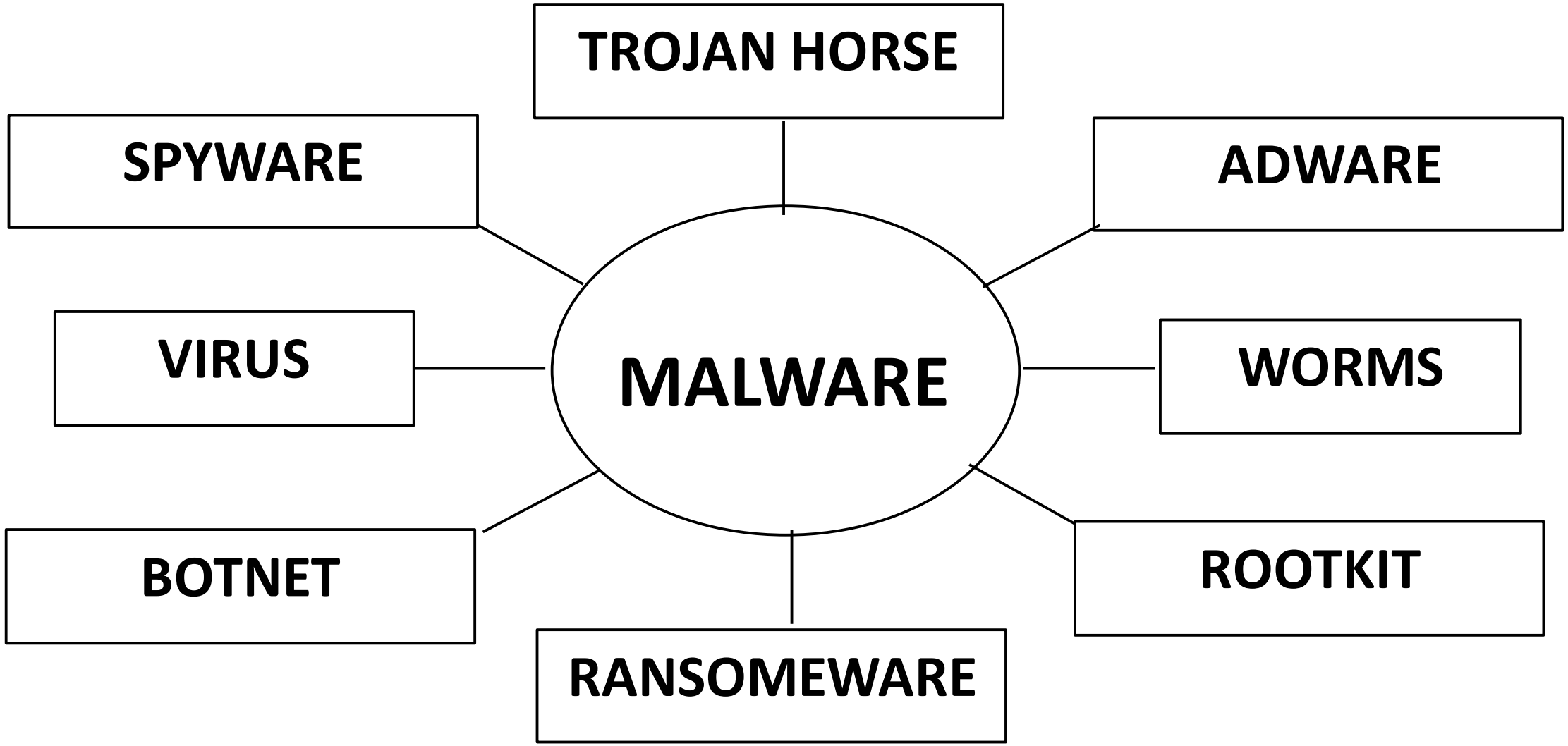
Lawrence Emmanuel

A SYSTEMATIC LITERATURE REVIEW OF MACHINE LEARNING FOR MALWARE: METHODS, ALGORITHMS, PERFORMANCE, LIMITATIONS AND FUTURE RESEARCH DIRECTION

INTRODUCTION

Malware short for malicious software, refers to various kinds of harmful software designed to disrupt, damage or gain unauthorized access to computer system

TYPES OF MALWARE



MALWARE DETECTION TECHNIQUES

Traditional Methods

Signature –Based Detection

- Uses a database of known malware signatures for identification.
- it is effective against known threats but
- fails with new or modified malware

Behavioural – Based Detection

- Analyse behaviour rather than code to detect malware,
- It identifies zero-day threats but
- have high false positives.

Non-Traditional Methods

Heuristic – Based Detection

- It utilizes machine learning to detect unknown malwares by analysing behaviour during runtime but
- it has a reduced false positives rate

MACHINE LEARNING IN MALWARE DETECTION (CONT'D)

- Machine learning is a subset of Artificial Intelligence (AI), which uses algorithms that learn from data to make predictions. These predictions can be generated through the different types.

TYPES OF MACHINE LEARNING

- ***Supervised***: Involves training an algorithm using a labeled dataset
- ***Unsupervised***: Involves training algorithms on unlabeled data to discover patterns or structures in the data.
- ***Semi-Supervised***: The algorithm is trained on both labeled and unlabeled data to leverage the labeled data for improving the learning process while taking advantage of the additional information provided by the unlabeled data.
- ***Reinforcement*** learning: learns to make decisions by interacting with its environment

Each ML type is suited for different types of data analysis tasks

PERFORMANCE METRICS

Performance metrics are essential tools used to evaluate the effectiveness of various models. These metrics provide insight into the quality of the classification process, as each metric presents a unique evaluation of the model. Some of these metrics are:

- **Accuracy:** This is the portion of the test set that the model predicts correctly
- **Precision:** It is the portion of the test set that the model predicts incorrectly
- **False Positive Rate:** it is the portion of the test set that the model predicts falsely as positive when it is negative
- **False Negative Rate:** it is the portion of the test set that the model predicts falsely as negative when it is positive

THE PREFERRED REPORTING ITEMS FOR SYSTEMATIC REVIEWS AND ANALYSES (PRISMA 2020), (credit: MOHER *et al*. 2009)

Consist of two phase

Phase 1: Planning

The planning phase involves identifying the study's objective and defining appropriate protocols to be followed while performing the review.

This review entailed the careful selection and evaluation of papers based on specific criteria

PHASE 2: IDENTIFICATION OF NEW STUDIES VIA DATABASE

Performing the review

Following the guidelines of the PRISMA 2020 framework, this phase is mainly accomplished through five sub-phases:

i) Record Identification

ii) Record Screening

iii) Report sought for retrieval

iv) Report eligibility assessment

v) Report inclusion.

FORMULATING RESEARCH QUESTIONS

RQ1 – What sources of dataset have been used in the research areas?

RQ2 – Which are the most frequently used model/algorithm in Machine learning to detect malware?

RQ3 – what are the most frequently used Performance Metrics to evaluate the effectiveness of various models?

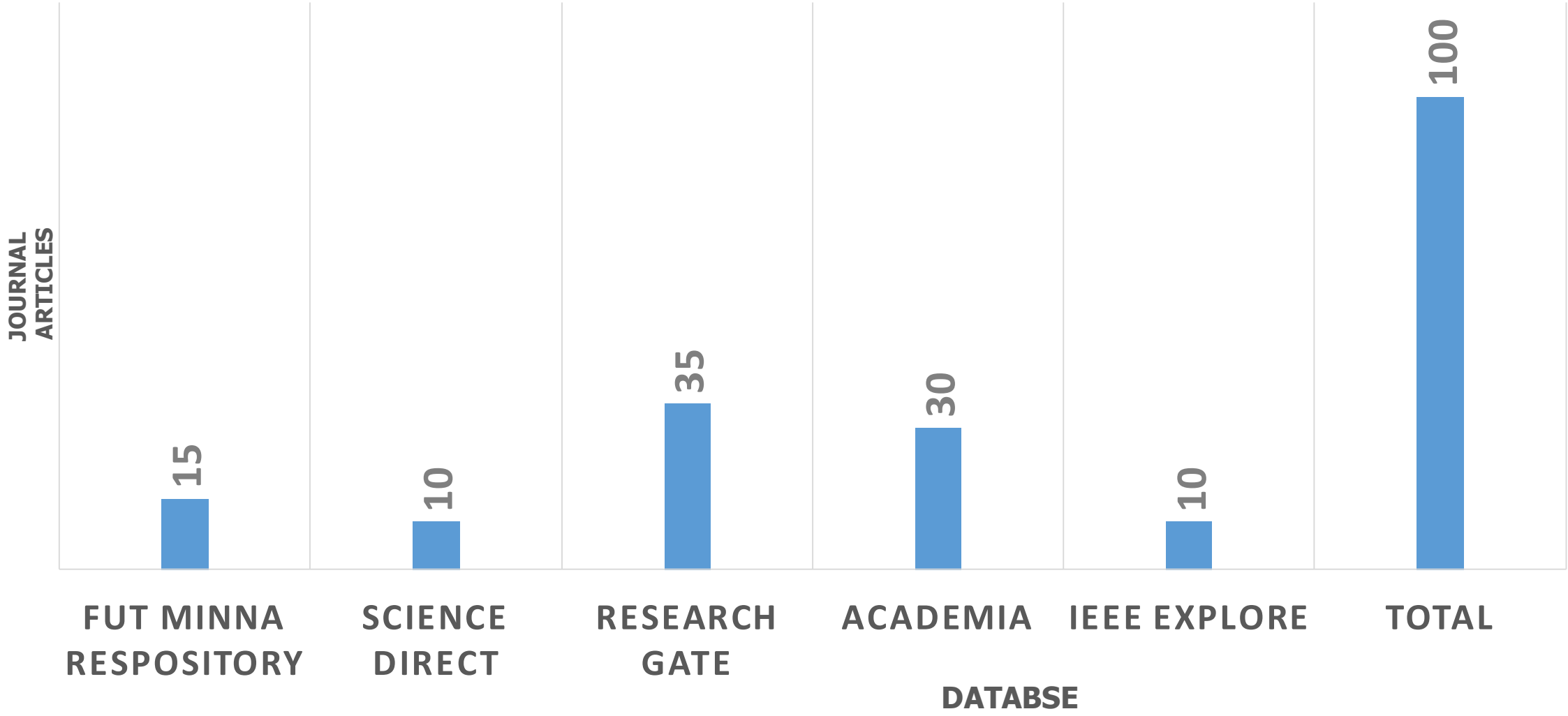
Defining Data Sources and Search Strategy

Records identified from databases

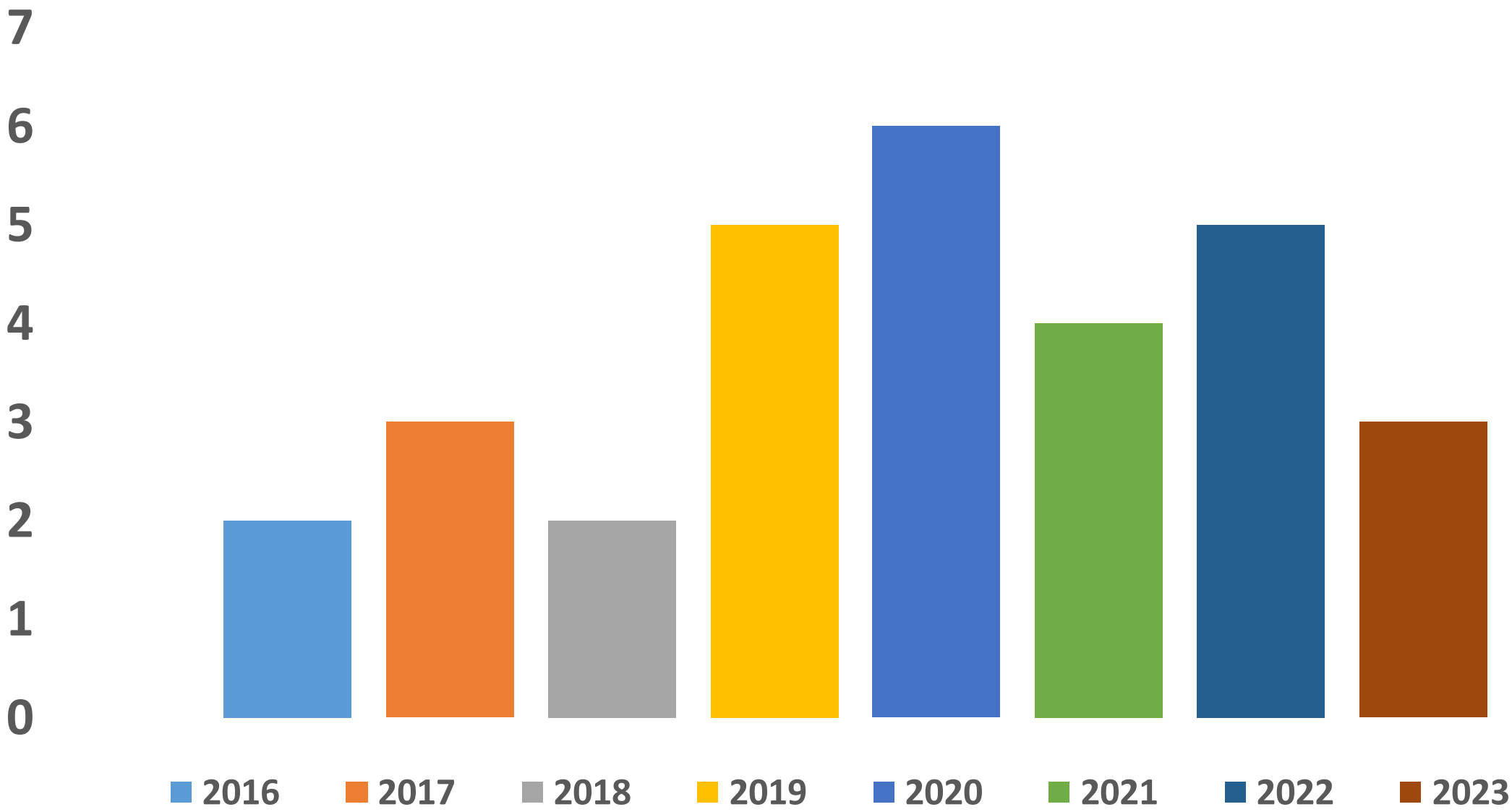
To obtain reliable data for the study we search some database, these databases include:

- Federal University of Technology, Minna (FUT Minna) Repository
- Science Direct
- IEEE Xplorer
- Academia and
- Research-Gate

Journal articles plotted against database



ARTICLES DISTRIBUTION PER YEAR



RECORDS SCREENING

Criteria	
Inclusion	A1. The full article was written in English A2. Contain Journal and conference article papers A3. the search terms are mentioned either in the title, abstract, or keywords A4. Solutions contain malware detection techniques using machine learning Algorithm
Exclusion	B1. The article was written outside the range 2016-2023. B2. Book and white paper B3. Duplicate copies indexed in other databases B4 Papers whose full text is not available B5 All papers that are not written in English were excluded B6. Literature review or overview of other paper B7. Papers not explicitly related to malware detection using machine learning

REPORT SOUGHT FOR RETRIEVAL

Quality assessment criteria

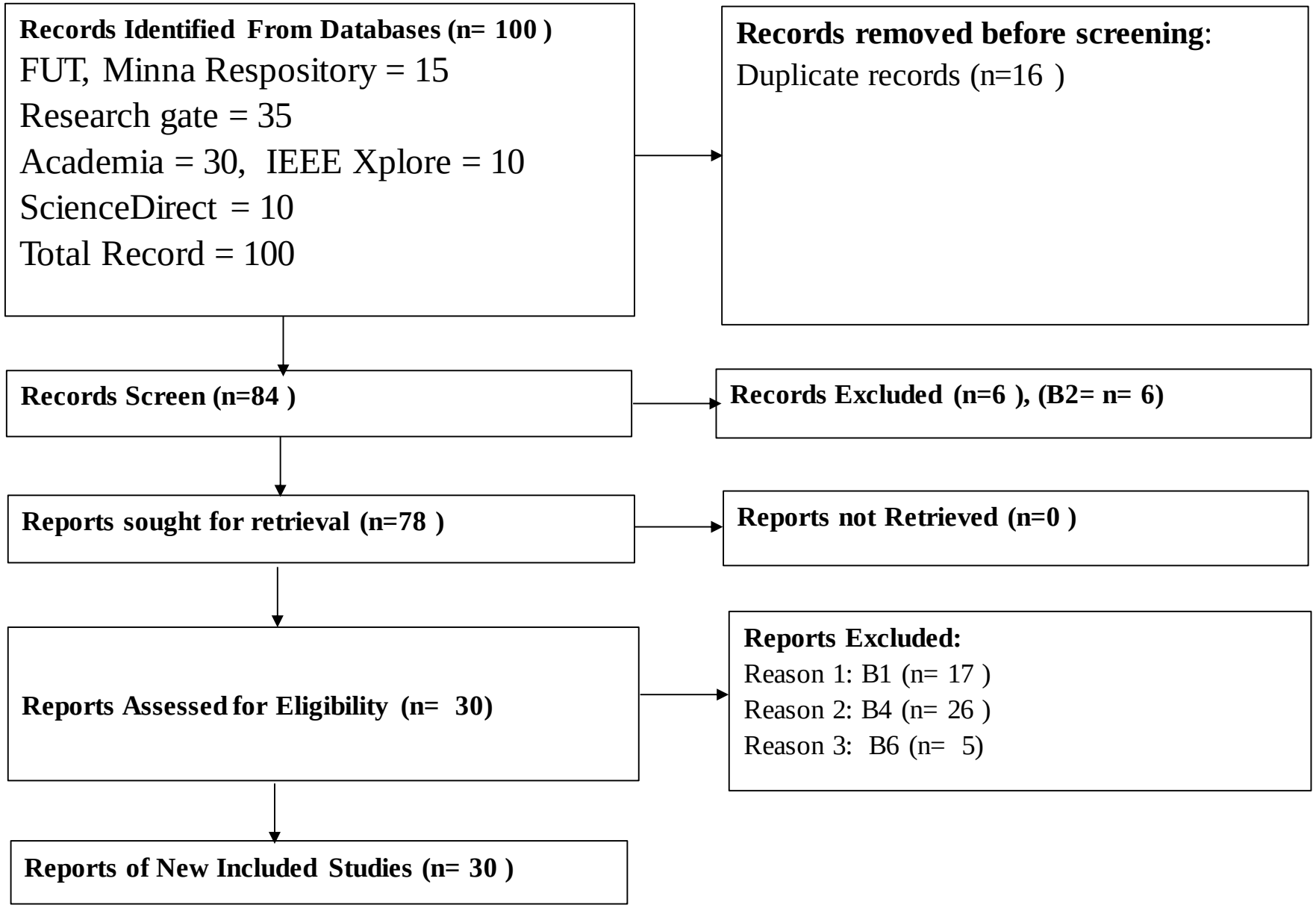
Number	Defined quality assessment criteria
1	Are the goals of the study clearly stated?
2	Did the study provide details of the dataset?
3	Are the experimental evaluations clearly presented and discussed?
4	Did the study clearly state the machine learning models used?
5	Did the study clearly state the performance metrics used?

REPORTS INCLUDED STAGE

This phase is commonly referred to as data extraction. For an article to be included in this phase, it must have the following characteristics:

- They were journals or conference papers
- Written in English language
- Published between 2016-2023, and
- Offered solutions on malware detection using machine learning.

IDENTIFICATION OF NEW STUDIES VIA DATABASE



COMPARATIVE ANALYSES OF ML MODELS

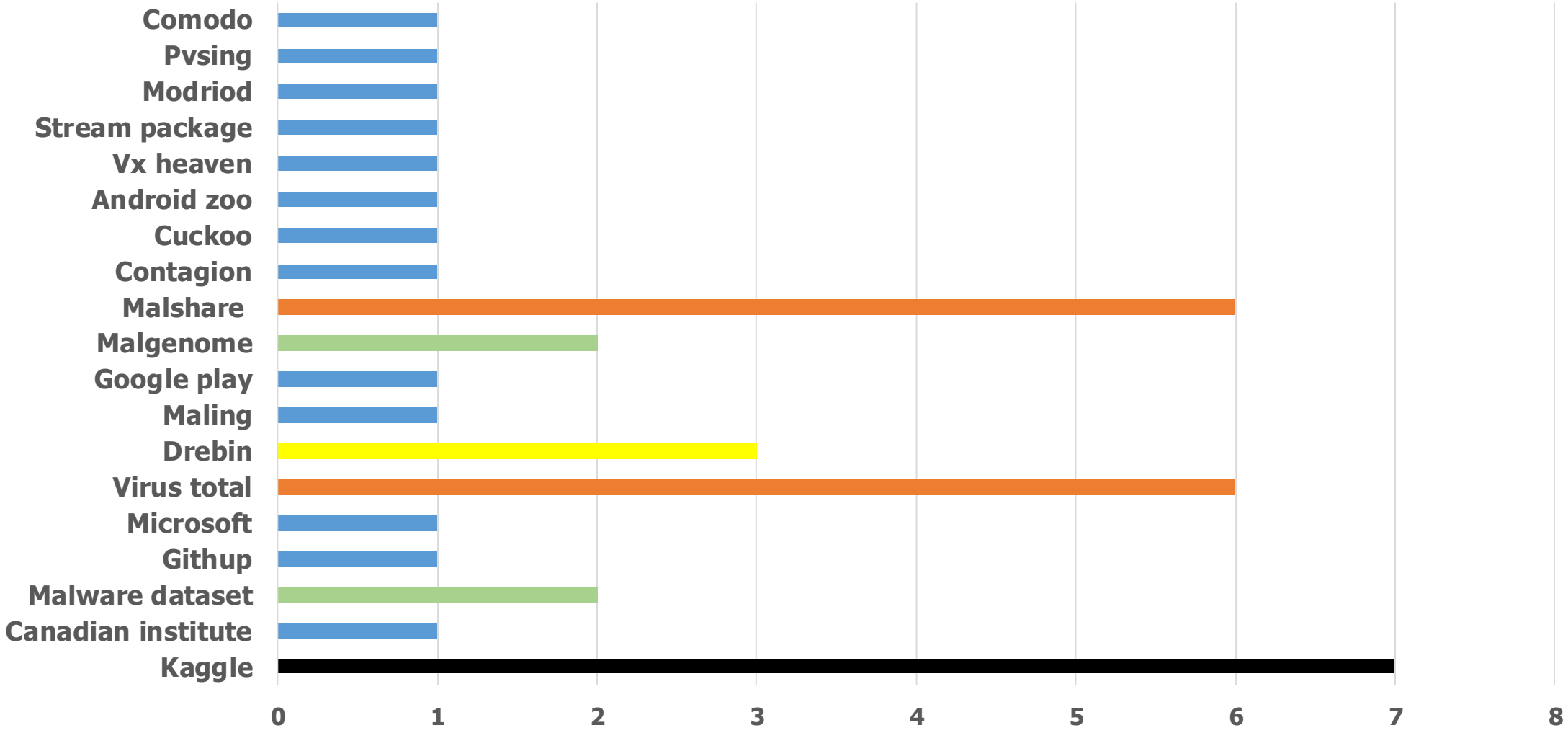
S/N	Authors	Title	Algorithm	Source of Dataset	Application	Metrics	Result
1	Reddy et al 2023	Behaviour based malware detection using machine learning	DT, RF , CNN	Kaggle library	windows	Accuracy, Precision, Recall , FPR, FNR	NA
2	Muhammad and Tao Feng 2023	Evaluation of machine learning algorithms for malware detection	KNN, DT, RF, AdaBoost, SGD, GNB	Kaggle library	windows	Accuracy, Precision, Recall F1 - score	RF had the highest accuracy of 100%
3	Mohd et al., 2023	Detecting Malware with Classification Machine Learning Techniques	Naïve Bayes, Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Decision Tree	Kaggle.	windows	FPR, FNR, TPR, TNR	Decision Tree and Random Forest display superior performance with accuracy of 99.27%

28	AeGuen *et al.,* 2017	A Multimodal Deep Learning Method for Android Malware Detection using Various Features	SVM, RF, DNN, and multimodal neural network (MNN)	VirusShare	Adriod	multimodal deep learning (MDL)	Not Mentioned
29	M. Sujithra, G. Padmavathi 2016	Enhanced Permission-based Based Malware Detection in Mobile Devices Using Optimized Random Forest Classifier with PSO-GA	Random Forest (RF), Classification and regression trees (CART), and J48	Virus Total	Andriod	TPR, FPR Precision-Recall and Accuracy	random forest of correctly identified instances with 86.8%.
30	William 2016	A Deep Learning Framework for Intelligent Malware Detection	ANN, SVM, DT, NB and DL4MD	Comodo Cloud Security Center	windows	TP, FP, TN, FN and Accuracy	DL4MD has the best accuracy of 95.6%

RQ1 – What source of dataset have been used the most in the research area?

The question can be answered by examining the dataset sources used by each research paper and documenting the findings. A total of 30 articles were used to address RQ1, from the graphical analyses that was shown to depict the various dataset sources , the grapgh illustrates that Kaggle.com was the most used dataset source by authors.

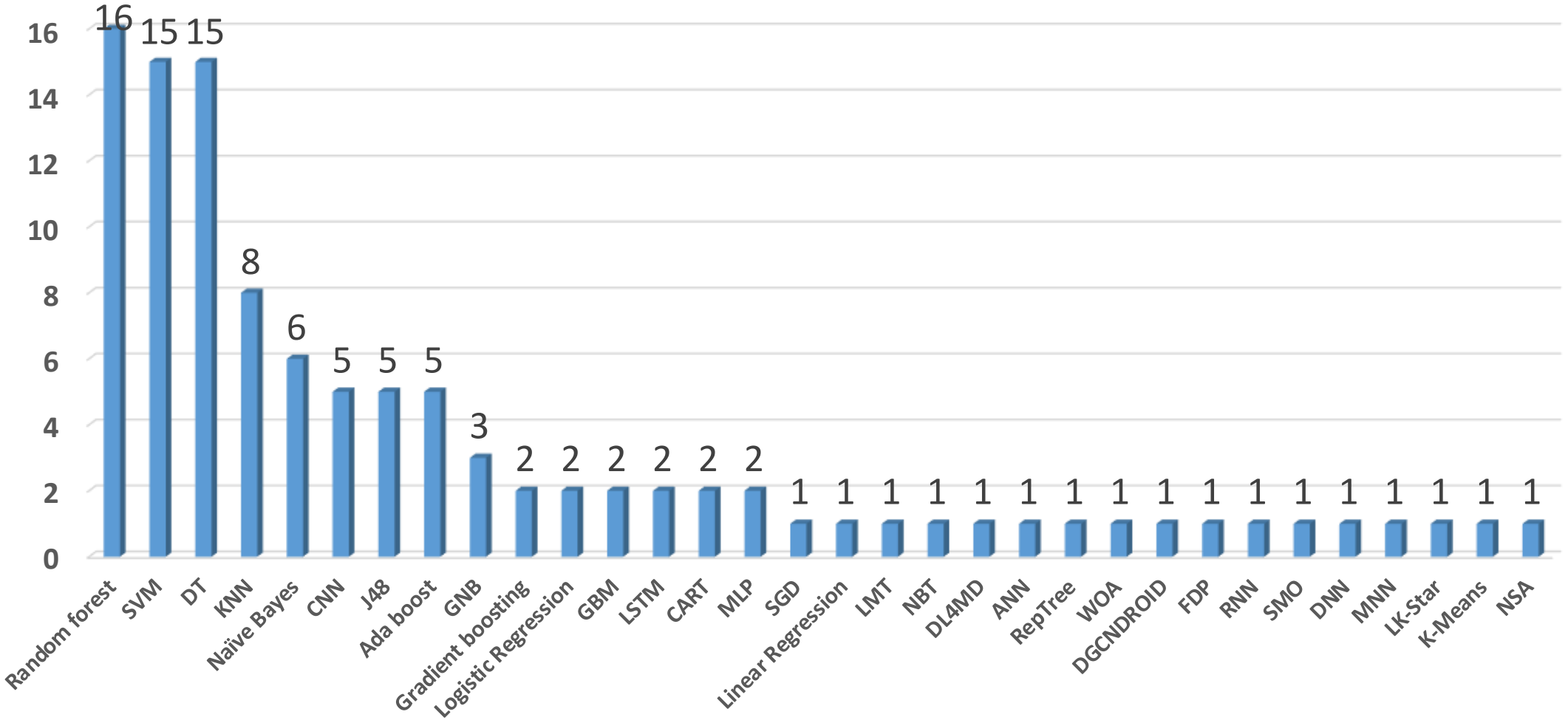
Source of Dataset plotted against Frequency of usage



RQ2 – Which is the most frequently used algorithm in Machine learning to detect malware?

The results presented in the table indicate that some researchers, such as Muhammed *et al*., (2021) and Eliel *et al*.,(2022) employed a single algorithm, specifically Convolutional Neural Networks (CNN), for malware detection. In contrast, other researchers, like Kakavand *et al*. (2018) and Joshi and Mahagaonkar (2022), utilize two distinct algorithms. Furthermore, the remaining researchers implement more than two different algorithms for this purpose. Among the algorithms employed, Random Forest (RF) is the most frequently used, appearing in 16 out of the 30 selected studies. Following RF, Support Vector Machine (SVM) and Decision Tree algorithms are used by 15 authors each. The various algorithms and their respective frequencies of usage

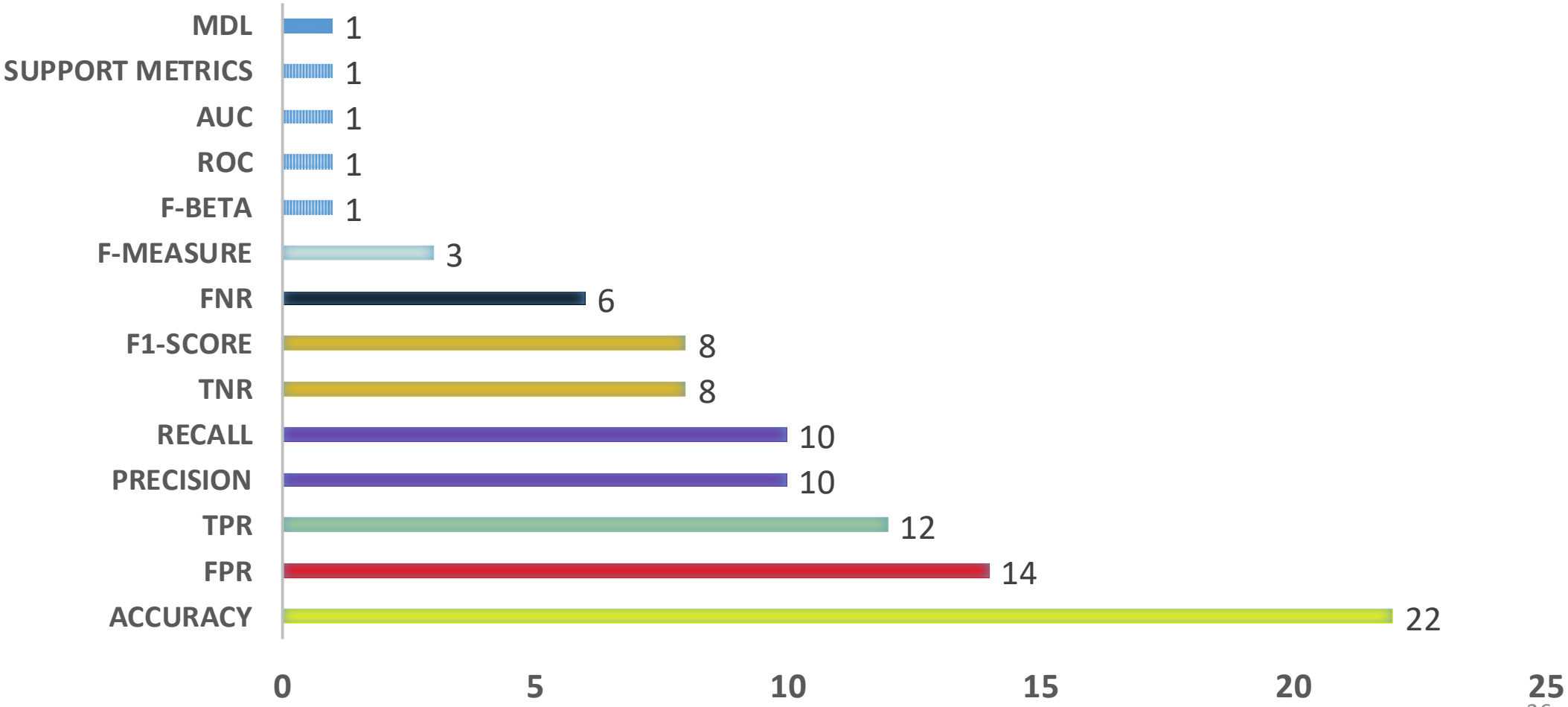
GRAPH OF ALGORITHMS USED IN THE INCLUSION ARTICLES AND THE FREQUENCIES OF USAGE



RQ3 – what are the most frequently used Performance Metrics to evaluate the effectiveness of various models?

The question is addressed by analyzing the performance metrics employed by different authors to evaluate each model. Accuracy metrics were the most commonly utilized, with 22 out of the 30 selected articles using this metric to assess model effectiveness. The performance metrics utilized in the selected articles were summarized graphically as

GRAPH OF PERFORMANCE METRICS AGAINST FREQUENCY



Discussion

An extensive SLR was performed on the recent literature about malware detection using machine learning techniques. The analysis provides a summary of the sources where the datasets were collected from, the various algorithms used, performance metrics employed, and the results obtained. The objective of the analysis was to examine the shift towards machine learning techniques for malware detection, which has recently gained popularity.

DISCUSSION CONT'D

The analysis of thirty (30) recent works revealed that supervised machine learning algorithms, such as Support Vector Machines (SVM), Random Forests, and Decision Trees, were predominantly used by various researchers for detecting malware. Among these algorithms, Random Forest was demonstrated to be an effective classifier in multiple cases.

LIMITATIONS OF THE SYSTEMATIC LITERATURE REVIEW

- ***Publication Bias***: Potential bias towards studies published in well-known academic databases may overlook relevant research from smaller venues.
- ***Temporal Scope***: Limited timeframe (2016-2023) exclude older but relevant studies on malware detection.
- ***Language Bias***: Exclusion of non-English articles overlooked valuable research published in other languages.
- ***Search Strategy***: Variations in indexing and keyword usage across databases result in missed relevant studies.
- ***Generalizability***: The findings only apply to the specific situations and studies included.

KEY FUTURE RESEARCH DIRECTIONS

- Advanced hybrid methods should be developed, combining signature-based, behaviour-based, and machine-learning techniques for more comprehensive malware detection.
- Methods for quickly finding new malware software should be explored, potentially utilizing reinforcement learning programs to assist with this.
- Methods can be developed to collect and analyze threat intelligence data in real time, allowing for proactive identification of new malware threats.

CONTRIBUTION TO KNOWLEDGE

This work contributes to knowledge by providing insights into the ongoing challenges and advancements in malware detection, emphasizing the need for more robust and accurate techniques to combat this persistent threat to computer systems.

CONCLUSION

The review emphasizes the importance of leveraging machine learning models to enhance malware detection capabilities, by shedding light on effective approaches for combating malware and safeguarding computer systems against malicious attacks. Hence, enabling cybersecurity professionals to stay ahead of emerging threats. By synthesizing insights from a wide range of research papers, this review provides valuable guidance for future research directions and the development of more robust malware detection systems.

THANKS FOR LISTENING