Noname manuscript No.
(will be inserted by the editor)

A Threat Monitoring System for Intelligent Data Analytics of Network Traffic

Lucas C. B. Guimarães · Gabriel Antonio F. Rebello · Gustavo F. Camilo · Lucas Airam C. de Souza · Otto Carlos M. B. Duarte

Received: date / Accepted: date

Abstract Cybernetic attacks have been increasingly common and cause great harm to people and organizations. Late detection of such attacks increases the possibility of irreparable damage, with high financial losses being a common occurrence. This article proposes TeMIA-NT (ThrEat Monitoring and Intelligent data Analytics of Network Traffic), a real-time flow analysis system that uses parallel flow processing. The main contributions of the TeMIA-NT are: i) the proposal of an architecture for real-time detection of network intrusions that supports high traffic rates, ii) the use of the structured streaming library, and iii) two modes of operation: offline and online. The offline operation mode allows evaluating the performance of multiple machine learning algorithms over a given dataset, including metrics such as accuracy, F1-score, and area under the curve (AUC). The proposed system uses dataframes and the structured streaming engine in online mode, which allows detection of threats in real-time and a quick reaction to attacks. To prevent or minimize the damage caused by security attacks, TeMIA-NT achieves flow-processing rates that reach 50 GB/s.

Keywords machine learning · big data · security · threat detection · stream processing

Lucas C. B. Guimarães
Grupo de Teleinformática e Automação
Univesidade Federal do Rio de Janeiro
Tel.: +55 21 3938-8635
E-mail: chagas@gta.ufrj.br

Otto Carlos M. B. Duarte
Grupo de Teleinformática e Automação
Univesidade Federal do Rio de Janeiro
Tel.: +55 21 3938-8640
E-mail: otto@gta.ufrj.br

1 Introduction

Cybercrime is one of the major challenges introduced by the exponential growth of the Internet. According to Cybersecurity Ventures [1], damages related to cyber-attacks are projected to reach US$6 trillion by 2021. Besides, the growth and popularization of areas such as Big Data and the Internet of Things pose even more significant challenges to cybersecurity. The introduction of billions of low power computing devices connected to the network increases the impact of possible attacks, as these devices can be easily hacked and compromised on a large scale [2,3,4]. The large volume of data to be analyzed in real-time also increases the complexity of classifying network traffic and detecting threats [5]. Finally, the average time to detect an attack is a crucial factor in the impact of cyber threats. More than a quarter of cyber-attacks take a long time before being discovered, with this time often ranging from weeks to months [6]. The late detection of an attack exponentially increases the risk of financial losses and the risk of irreparable damage. The long time is due to the need for human intervention in these situations, significantly affecting the efficiency of dealing with threats.

In this scenario in which security is a fundamental aspect, the need for systems capable of guaranteeing safe and reliable network use is increasing. Solutions based on Security Information and Event Management (SIEM) tools partially mitigate the problem by providing real-time network monitoring. This type of solution, however, is still highly dependent on the intervention of experts and is based on threat signature databases, therefore being inefficient in the detection of new attacks. Using machine learning algorithms for threat detection, on the other hand, automates the detection process and meets the required agility to pre-

vent and mitigate network attacks. It is of utmost importance to select algorithms that perform well in the classification process, without harming accuracy and other evaluation metrics. Previously, our research group (GTA/UFRJ) proposed CATRACA [7], a tool that uses machine learning to detect threats in real-time.

This paper proposes TeMIA-NT: Threat Monitoring and Intelligent Data Analytics of Network Traffic, an intelligent threat monitoring and detection system based on machine learning and distributed processing in clusters. TeMIA-NT proposes and develops an entirely new distributed processing system with significant improvements in machine learning processing optimization. Our proposal focuses on the intelligence, scalability, and performance required to process large volumes of data while optimizing multiple machine learning algorithms to meet the diversity of new attacks. To increase performance, TeMIA-NT implements distributed processing entirely in Scala language and uses dataframe structures, instead of the standard Resilient Distributed Datasets (RDD) structure on the open-source Apache Spark platform. TeMIA-NT offers many options for machine learning algorithms and the possibility of optimizing hyperparameters, allowing testing, selecting, and adjusting the parameters of the best algorithm for each type of scenario. We implement the offline threat detection using the structured streaming library, which allows flow processing in micro-batches, with fault tolerance and reduced intervals. Online threat detection uses the continuous processing mode, which enables our proposal to perform similar to a native stream processing tool.

The rest of the article is organized as follows. Section 2 presents papers with themes related to the article. Section 3 introduces the Apache Spark platform, as well as its data structures and its machine learning library. Section 4 presents the machine learning algorithms used during the performance analysis, as well as a brief look at their hyperparameters. Section 5 offers a detailed look at the network traffic dataset used to test the proposed system, while Section 6 presents the system's architecture and features. Section 7 presents and analyzes the performance tests and their results, and Section 8 presents the author's final considerations and concludes the work.

2 Related Works

New challenges in the intrusion detection area arise due to the high volume of traffic, a large number of IoT devices, distributed denial of service attacks, and zero-day attacks [8, 9, 10]. To meet these challenges, the use of machine learning techniques to classify flows in real-time became popular [7, 11, 12, 13]. The classification of large volumes of data at high speeds available employs three main distributed processing platforms: Apache Spark, Apache Storm, and Apache Flink. The fundamental difference between the platforms is that Spark performs batch processing while the Storm and Flink platforms perform native flow processing.

The Open Security Operations Center (OpenSOC) [14] is an analytical security framework for monitoring large amounts of data. OpenSOC originated a new project, Apache Metron [15], that is a tool that comprises the acquisition of different types of data, distributed processing, enrichment, storage, and visualization of results. Metron allows the correlation of security events from various sources, such as logs of applications and network packages. For this purpose, the framework uses distributed data sources, such as sensors on the network, logs of security element events, and enriched data called telemetry sources. The tool also provides a historical base of Cisco network threats.

Based on the Apache Spark Platform [16], there are the Apache Spot, Stream4Flow [17], and Hogzilla. Apache Spot is a project still in the incubation stage that uses telemetry and machine learning techniques for analyzing packages to detect threats. The Stream4Flow prototype uses the Elastic stack to view network parameters, however, it lacks the intelligence to perform anomaly detection. The Hogzilla tool provides support for Snort, SFlows, GrayLog, Apache Spark, HBase, and libnDPI, offering network anomaly detection. Hogzilla also allows visualizing network traffic, using Snort to capture packets, and obtaining features through deep packet inspection. Stream4Flow captures packets using IPFIXcol and only considers header information. In our work, we use the flowtbag software, which captures various flow statistics. In addition, our offline processing mode allows updates to the machine learning model, further promoting the detection of new threats.

We select the Apache Spark platform to develop the TeMIA-NT because it is the most adopted among the examined Big Data processing platforms. Spark offers more possibilities for machine learning algorithms and is the one with the largest active community. Nevertheless, to the best of our knowledge, TeMIA-NT is the only available system to use the recent structured streaming technology in batch and continuous modes in Apache Spark, allowing the selection among several machine learning algorithms, and operating in offline and online modes.

3 The Apache Spark Platform

We use Apache Spark [16], a distributed processing platform for Big Data, providing an interface for programming in clusters with parallelism and fault tolerance, to develop the system in this paper. We chose the Spark platform due to its efficiency, its great acceptance in the market, and because it has a wide library of machine learning algorithms. The platform also supports multiple programming languages, including Python, Scala, R, and Java.

The main feature of Apache Spark is how it processes data: all operations that involve reading and writing intermediate results are done in memory. Spark is efficient for applications that perform multiple data transformation iterations in a distributed environment, avoiding time-consuming disk operations [18].

The Spark platform provides several libraries, such as Spark Streaming for real-time flow processing and GraphX for parallel graph computing. Also, Spark provides MLlib, a library that implements parallelizable and efficient machine learning algorithms in a distributed environment, making the platform an option for classifying network traffic. We use algorithms from the MLlib library to do performance analysis, which creates the machine learning models used for traffic classification.

3.1 Data Structures

Because of the growing impact of Big Data, Zaharia *et al.* designed and developed Apache Spark to provide enterprise-level distributed processing for large datasets [16]. The data structures used by Spark play an essential role in the fast and efficient data processing, being responsible for its organization, management, and storage; they also provide functions and operations to make more efficient data processing.

3.1.1 Resilient Distributed Datasets

The first Spark data structure developed for distributed processing was resilient and distributed datasets (RDD). This structure is an immutable and, therefore, resilient dataset, partitioned in the cluster nodes. It can be operated by a low-level API, offering multiple transformations and functions. A crucial feature of this data structure is to provide computing resources in memory, providing the agility observed in Spark operations. Another essential feature is the use of lazy evaluation, which computes expressions or functions only when their results are needed, optimizing the execution time by avoiding unnecessary calculations. RDD also offers fault tolerance: each RDD can reconstruct lost data automatically, based on data within other nodes in the cluster. Since RDDs are immutable, they can be created or retrieved at any time, making data sharing and replication a simple process.

3.1.2 DataFrame & Dataset

DataFrames and Datasets are the other data structures implemented by Apache Spark. These structures differ from RDD in that they are structured as tables in a relational database: RDDs do not specify rows and columns, thus queries in RDDs with a large number of records require longer periods to complete. On the other hand, DataFrames and Datasets follow a schema, which lists the columns and the information they contain. As the data implemented through DataFrames and Datasets are structured, Spark implements performance optimizations in terms of processing time and memory consumption through the Tungsten [19] and Catalyst Optimizer [20] projects.

These data structures act similarly, differing only in terms of type handling: Datasets implement a strongly typed API, while DataFrames implement an untyped API. An untyped API allows parsing errors to go unnoticed during compilation time. Differently, a strongly-typed API detects these errors at compile-time, reducing the possibility of errors occurring during the execution of the program. Since Python and R do not have compile-time type security, these languages only implement DataFrames.

3.2 MLlib Library

The purpose of the MLlib [21] library is to allow the use of machine learning techniques on the Apache Spark platform, implementing them in an efficient and scalable way through a high-level API. These techniques include standard classification, regression, clustering, and collaborative filtering machine learning algorithms, such as decision tree, linear regression, k-means, alternating least square, among others.

The library also offers featurization methods, allowing the Apache Spark platform to carry out the preprocessing of datasets before machine learning methods are applied. The application includes techniques that reduce dimensionality and rely on both the selection and extraction of features. These methods also allow the transformation of those features, such as normalization.

MLlib also provides multiple utilities to facilitate data processing, including statistical methods used to obtain results in terms of evaluation metrics, such as

accuracy and AUC, and linear algebra methods. There are also methods responsible for optimizing the execution pipelines, allowing algorithms and models to be saved and loaded from memory as necessary.

4 Machine Learning and Hyperparameters

Since they possess different logics and assume different characteristics of the input data, machine learning algorithms present different results depending on the target problem. Therefore, it is important to evaluate which algorithms offer the best results for analysis and classification of network traffic. As such, the following algorithms made available through the MLlib library were implemented in the proposed system, so that the user can verify which offers the best solution to their dataset: Naïve Bayes, Logistic Regression, Support Vector Machine, Multilayer Perceptron, Decision Tree, Random Forest and Gradient-boosted Tree.

Hyperparameter tuning is the optimization of machine learning models, obtaining the best set of hyperparameters of an algorithm for a given dataset. Hyperparameters are the parameters that determine the learning process, and thus are selected prior to the training, in contrast with regular parameters that are learned during the training such as weights and bias. It's an exhaustive process, since it requires multiple executions of the learning algorithm, each time changing a specific parameter.

Some hyperparameters are unique to a given algorithm, such as Naïve Bayes's smoothing. However, it is common for algorithms to share a number of hyperparameters; this can be seen on Decision Tree-based ones, all of which include an hyperparameter for setting the maximum tree depth, as well as on iterative ones, which all include hyperparameters for the maximum number of iterations and the minimum threshold necessary for convergence.

Naïve Bayes: The naïve Bayes methods are a set of probabilistic classifiers that work through the application of Bayes' theorem. This theorem, given by

$$P(c|x) = P(c)\frac{P(x|c)}{P(x)} \tag{1}$$

indicates the probability that an event c will occur knowing that a given event x has happened. The parameters used by the equation are the *a priori* probabilities of c and x, as well as their likelihood. Since the algorithm performs the classification through a simple mathematical calculation, resulting in linear execution time, it is easily scalable for large datasets and several features. However, the accuracy obtained by this classifier may be lower than that obtained by other algorithms, since it assumes that analyzed elements are statistically independent, which may not be valid depending on the chosen dataset.

The implementation of this algorithm on the Apache Spark platform provides two hyperparameters for tuning: the model type and the smoothing value. The model types available are: multinomial, complement, Bernoulli and Gaussian. Multinomial models are commonly applied to datasets containing categorical data. Complement is an adaptation of the multinomial method, used to better deal with unbalanced datasets. Bernoulli assumes that the data follows a Bernoulli distribution; as such, each feature must have binary or boolean values. Gaussian models assume that the probability distribution of the records follows a Gaussian distribution, allowing the models to handle continuous data. In turn, the smoothing hyperparameter is used to handle record values not observed during model training. Finding a new value results in a probability of zero and, as all probabilities are multiplied in the Bayesian equation, the final probability is also zero. Thus, the objective of the smoothing parameter is to ensure that the probability of each record is always greater than zero. Setting smoothing to 1 implements the Laplace smoothing, which is used by default by Apache Spark. Setting the smoothing parameter to values less than 1, but greater than 0 implements what is known as the Lidstone smoothing.

Logistic Regression: Logistic regression is a statistical algorithm that seeks to model the probability of a phenomenon, using the *sigmoid* function as a discriminant function. The function curve generated returns the likelihood of the data to be positive or negative. Based on this, the algorithm estimates the probability of new inputs to be or not in a certain class, making the binary classification.

The hyperparameters provided by Apache Spark are: elastic net parameter, regularization parameter, maximum number of iterations, convergence tolerance, threshold, fit intercept and standardization.

The elastic net and regularization parameters influence the regularization applied during the calculation of the algorithm. The purpose of regularization is to reduce the overfitting of the model, adding a penalty to the loss function. Setting the elastic net parameter to 0 results in the use of the $L2$ norm as a penalty, while setting this value to 1 results in the $L1$ norm; intermediate values result in the proportional application of both norms. The $L1$ norm is calculated by adding the absolute values for each feature, while the $L2$ norm is obtained by adding these values squared. The penalty is then multiplied by the regularization parameter: small

values for this parameter can still result in overfitting, while values that are too large may result in underfitting the model. The maximum number of iterations and the convergence tolerance define stop conditions for the execution of the algorithm. The convergence tolerance defines that the execution of the algorithm must be interrupted if the improvement between two iterations is less than the defined tolerance; the maximum number of iterations interrupts the algorithm if the tolerance is not reached after a certain number of iterations. The threshold defines the value that is used to classify the records as belonging to a certain class, being a value between 0 and 1. The fit intercept is a Boolean hyperparameter, which defines whether a constant should be added to the decision function. Finally, standardization is another Boolean hyperparameter, which defines whether the training features are standardized by the algorithm itself. It is essential that a standardization method be applied to the dataset if regularization is used, as regularization is significantly influenced by the values of the features of the training set [22]. This hyperparameter defines that the standardization is done by the algorithm itself, however it can be set as $False$ if this step is performed during the preprocessing of the data.

Support Vector Machine: The support vector machine (SVM) algorithm maps the training data in space and performs a binary classification defining a hyperplane, the decision boundary. This hyperplane is set to partition the space, aiming to maximize the separation margin between the closest points of each of the classes. Altogether a larger margin results in a better generalization of the model.

A fundamental aspect of this method is the definition of its kernel function, responsible for the mapping done in the feature space. There are several kernel functions, the most used of which are: radial base function (RBF), polynomial, hyperbolic tangent and sigmoid. However, the SVM algorithm used by Spark presents the linear kernel as the only available option. Similar to logistic regression, the stop conditions of the algorithm are determined by the maximum number of iterations and the convergence tolerance. Also present is the hyperparameter that determines whether the training features are standardized in the preprocessing, or during training, as well whether the fit intercept is used. Finally, the regularization parameter is also available, acting on the impact of the penalty; however, unlike SVM, logistic regression supports only the $L1$ norm.

Multilayer Perceptron: Multilayer perceptron is a neural network model that works by employing multiple perceptrons, which act as the network's "neurons", who are delegated the tasks of performing small calculations and forwarding their results to other perceptrons. The perceptrons are organized in layers, with each perceptron in one layer being fully connected to the perceptrons in the next layer. The first layer receives the input features from the dataset, while the last layer represents the classification results.

Each perceptron uses an activation function to connect with others, based on the results of the previous layers and the adjusted weights for each output connection. These activation functions are non-linear, allowing the acquisition of non-linear models, but increasing the time required to obtain the model. One of the most used activation functions is the logistic (or sigmoid) function.

Another technique used by the multilayer perceptron is backpropagation; this algorithm works by calculating the gradient of the loss function concerning each weight by the chain rule, iterating one layer at a time from the last layer to avoid redundant calculations of intermediate terms. In this way, it is possible to update the weights of each layer to minimize losses.

The main hyperparameters that must be defined when using the multilayer perceptron are the number of hidden layers and the number of neurons in each layer. Most classification problems can be solved efficiently with one or two layers, while using many layers tends to result in considerably longer processing times for ever lower returns. For the number of neurons in each layer, a commonly adopted method is to use a single hidden layer, with the number of neurons being equal to the average between the number of characteristics and the number of labels in the data set. Another relevant hyperparameter is the learning rate, also known as the step size. This value acts in updating the weights during the execution of the algorithm; using a very low learning rate can result in long run times and overfitting, while higher learning rates can result in models wit lesser performance. Other hyperparameters provided by Apache Spark for this algorithm include the maximum number of iterations, as well as the minimum tolerance for optimization, similar to SVM and logistic regression.

Decision Tree: The decision tree algorithm builds a tree in which each internal node evaluates a data feature. Each branch represents a decision around a possible value for the selected feature, and each final node in a branch indicates the class the element is most likely to belong to. Thus, the algorithm traverses the tree branches and evaluates the features of each node to estimate the sample probability to belong to a particular class. A great advantage of the decision tree algorithm is its ease of understanding and interpretation, being composed exclusively by rules in the "if-then-else" format.

The most important hyperparameters provided by Apache Spark for the decision tree are the maximum depth of the tree, the minimum gain of information, the minimum number of instances per node and the metric selected to calculate the impurity of each feature. The maximum depth controls the generalizability of the algorithm and directly influences the training and test time of the algorithm. Extremely deep trees tend to divide the entire training set into their correct labels, and as a consequence they are overfitted, while trees with few levels are unable to capture the variance present in the dataset and tend to have low classification performance. The minimum information gain hyperparameter is the value that must be obtained in order to consider the division of a given node, controlling the growth of the tree by restricting which nodes can be created to divide the dataset. The minimum number of instances per node controls the growth of the tree and determines the minimum number of samples required in the children of the node to generate the branch. The node of the tree that is not able to generate the minimum number of samples for the right and the left child becomes a leaf of the tree. A larger minimum number of samples can positively influence the model's accuracy for large datasets, since a low number can lead the model to behave randomly. Impurity measures the diversity of children raised using characteristics that meet the criteria for division. Thus, impurity is a criterion for selecting among all candidates a feature with greater diversity to perform the division of the node. Impurity is measured using the Gini index or entropy, the main difference of which being the slower computation of entropy.

Random Forest: The random forest is an ensemble learning algorithm, proposed by Breiman [23], that works by creating multiple decision trees. Breiman also proposed the bagging method to create different decision tree structures and capture distinct behaviors of a dataset.

The bagging method comprises two phases: bootstrap and aggregating [24]. The bootstrap phase consists of generating equally-sized datasets from the original training dataset through random sampling. Then, the method trains decision tree models from each sampled dataset. The goal is to build learning models with different structures that present different views when classifying new samples. In the aggregating phase, the method uses all different model structures of each local model to discover the correct class of a new sample. Each machine learning model classifies the sample, and the final result is the statistical mode of all classifications. This way, the method can generalize the behavior of new samples while minimizing variance.

In a random forest [23] with H trees, the predicted class $\hat{y}$ of a sample x is given by:

$$\hat{y} = f(x) = arg\max_{y \in Y} \sum_{j=1}^{H} I(y = h_j(x)), \tag{2}$$

where $h_j(x)$ returns the predicted class of x by tree h_j. The term $I(.)$ is the indicator function. The set Y represents the existing classes, which in our work are binary: 0 for normal flows and 1 for malicious flows.

The number of trees in the forest and the subsampling rate are the adjustable hyperparameters for random forests, in addition to the hyperparameters of decision tree models. As the number of trees grows in the forest, the classifier's performance increases due to the high variance of the built decision trees. However, after approximately 100 trees the metrics remain statistically equal, only increasing the processing time [25]. The subsampling rate hyperparameter specifies the size of the dataset used to train each tree in the forest, and is defined as a fraction of the size of the original dataset.

This algorithm usually presents better results than those obtained by working with only one decision tree, in addition to offering less risk of overfitting, but it has a considerably longer processing time. However, random forests are extremely parallelizable, since the training and classification of a single tree are independent of the set. The adoption of parallel processing reduces the complexity of the algorithm [26].

Gradient-boosted Tree: As well as the random forest algorithm, the gradient-boosted tree is an ensemble learning algorithm based on decision tree models. Unlike random forest, where each tree is trained individually,the gradient-boosted tree trains all trees iteratively, where the new trees uses the prediction of previous trees to offer a more accurate model.

The gradient-boosted tree uses the Boosting method to optimize the model after each iteration. Several shallow trees are created, calculating the loss based on that tree created. In the next iteration, the algorithm creates a tree that aims to reduce the loss value generated by the previous function. This process is interrupted if the algorithm reaches a stop condition, such as the maximum number of trees created or if the next tree does not improve the model's metrics.

This algorithm also has a high processing time and is not ideal for large datasets. As it is a tree-based algorithm, it presents the same set of parameters and hyperparameters as the decision tree and random forest, except for the learning rate, also known as step size hyperparameter. The learning rate is the hyperparameter that controls how complex the next tree built will be, giving more relevance to the mistakes done by the previous tree.

4.1 Structured Streaming

The real-time processing on the Apache Spark platform was initially implemented through the Spark Streaming library, which allows continuous processing of RDDs through the DStream API. With the introduction of DataFrame and Dataset as new data structures, the Structured Streaming library was developed to handle these structures in real-time while maintaining the optimizations they introduced.

Structured Streaming allows the programmer to program in a similar way to the one in batch data processing, with the platform dealing with the implementation of specific flow processing techniques through a high-level API. Structured Streaming implements the micro-batch technique, with data received within a certain time interval being added to a batch to be processed; after processing, the result is added to a table, and the elements of the processed batch are discarded. Other advantages of the library include "exactly once" fault tolerance, as well as end-to-end latency of up to 100 milliseconds.

Another processing method provided by the library is the continuous processing mode. This mode allows latency as low as one millisecond but does not offer all the functions of the main library, supporting only projection and selection operations. It also has "at least once" fault tolerance, leaving aside the advantages of tolerating exactly once of the other processing method.

5 Dataset and Schema

A crucial aspect of the development of an intrusion detection system (IDS) is the need to check its performance before it goes into operation. Thus, a dataset is used that contains both legitimate and malicious traffic. The most commonly used dataset in IDS development is the NSL-KDD [27], with other important datasets being the DARPA98 and the DARPA99. However, these and other datasets are often not recent, and in addition to using synthetic attack patterns and threats, may not portray the features of current network traffic.

The dataset used was obtained from traffic from a telecommunications operator [28], converted into flows using the flowtbag tool. Each flow is a sequence of packets, within a time window, which has certain features in common. The features used to group packets in flows were the 5-tuple (source IP, destination IP, source port, destination port, protocol), set commonly used in traffic analysis works. Grouping packages into flows, the flowtbag tool extracts 40 features for the construction of the data schema, including the number of packages sent and received, the minimum and maximum sizes of a package, among others. The complete list of features is presented on Table 1.

In the preprocessing stage, we identified that the dataset had seven features which contained only null values; after these columns were removed, we calculated the Pearson correlation matrix shown in Figure 1 to allow a better understanding of the remaining features. Through this matrix, it is possible to identify that multiple features possess high correlation between them. For instance, features 7, 8, and 9 have a high correlation with feature 6, since all these features are related to the flow size; features 26, 27, 28, and 29 also have a high correlation, with these four features representing time measurements of the flow.

The dataset used groups together a series of attacks common on computer networks, such as attacks focused on the application, transport and network layers. Most of the observed attacks occurred at the application layer, because even though each layer presents its vulnerabilities, the application layer allows less sophisticated attacks to occur, which can be performed by inexperienced attackers. The types of attacks presented below bring together the main attacks observed in the dataset used.

5.1 Application Layer

Attacks at this layer target some protocols such as DNS, HTTP, IMAP and Telnet. Injection of SQL code, transmission of malware and cross site scripting were some attacks observed.

SQL code injection: Web pages often make SQL queries to check the database for a user's credentials. If the site does not verify that the data entry corresponds to the expected one, the text boxes for login and password can serve as input for actions in the database. Malicious users can alter, add or remove information in the database, compromising the integrity of the system. This type of malicious request within a database is configured as an SQL code injection attack, where the attacker uses text boxes to insert code snippets directly into the database. The objective is to access sensitive information and obtain advantages over this information, such as bank passwords, credit cards, among others.

Malware: Malware is a general term for any type of malicious software. They can take partial control of a device by running other scripts in the background. They are usually developed by teams of malicious entities seeking profit from the proliferation of malware or its auction on the Deep Web. There are several types of malware, such as Adware, Spyware, Ransomware, Viruses, Trojans and Worms, but always with the same

Table 1: Meaning of all network dataset rows generated by flowtbag [29].

Category	Number	Name	Description
Identifier	1	srcip	Source ip address
Identifier	2	srcport	Source port number
Identifier	3	dstip	Destination ip address
Identifier	4	dstport	Destination port number
Identifier	5	protol	Application protocol TCP or UDP
Feature	6	total_fpackets	Total packets in the forward direction
Feature	7	total_fvolume	Total bytes in the forward direction
Feature	8	total_bpackets	Total packets in the backward direction
Feature	9	total_bvolume	Total bytes in the backward direction
Feature	10	min_fpktl	Size of the smallest forward packet
Feature	11	mean_fpktl	Mean size of forward packets
Feature	12	max_fpktl	Size of the largest forward packet
Feature	13	std_fpktl	Standard deviation from the mean of the forward packets
Feature	14	min_bpktl	Size of the smallest backward packet
Feature	15	mean_bpktl	Mean size of backward packets
Feature	16	max_bpktl	Size of the largest backward packet
Feature	17	std_bpktl	Standard deviation from the mean of the backward packets
Feature	18	min_fiat	Minimum amount of time between two forward packets
Feature	19	mean_fiat	Mean amount of time between two forward packets
Feature	20	max_fiat	Maximum amount of time between two forward packets
Feature	21	std_fiat	Standard deviation from the mean time between two forward packets
Feature	22	min_biat	Minimum amount of time between two backward packets
Feature	23	mean_biat	Mean amount of time between two backward packets
Feature	24	max_biat	Maximum amount of time between two backward packets
Feature	25	std_biat	Standard deviation from the mean time between two backward packets
Feature	26	duration	Duration of the flow
Feature	27	min_active	Minimum time that the flow was active before idle
Feature	28	mean_active	Mean time that the flow was active before idle
Feature	29	max_active	Maximum time that the flow was active before idle
Feature	30	std_active	Standard deviation from the mean time that the flow was active before idle
Feature	31	min_idle	Minimum time a flow was idle before becoming active
Feature	32	mean_idle	Mean time a flow was idle before becoming active
Feature	33	max_idle	Maximum time a flow was idle before becoming active
Feature	34	std_idle	Standard devation from the mean time a flow was idle before turn active
Feature	35	sflow_fpackets	Average number of packets in a forward sub flow
Feature	36	sflow_fbytes	Average number of bytes in a forward sub flow
Feature	37	sflow_bpackets	Average number of packets in a backward sub flow
Feature	38	sflow_bbytes	Average number of packets in a backward sub flow
Feature	39	fpsh_cnt	Number of PSH flags in forward packets
Feature	40	bpsh_cnt	Number of PSH flags in backward packets
Feature	41	furg_cnt	Number of URG flags in forward packets
Feature	42	burg_cnt	Number URG flags in backward packets
Feature	43	total_fhlen	Total bytes used for headers in the forward direction
Feature	44	total_bhlen	Total bytes used for headers in the backward direction
Feature	45	dscp	First set DSCP field for the flow
Label	46	class	Flow class, benign or malicious

objective of extracting information to obtain advantages on the infected machine.

Adware: Intensive advertising software, displaying ads based on information collected about the victim. They can serve as an entry for other malware.

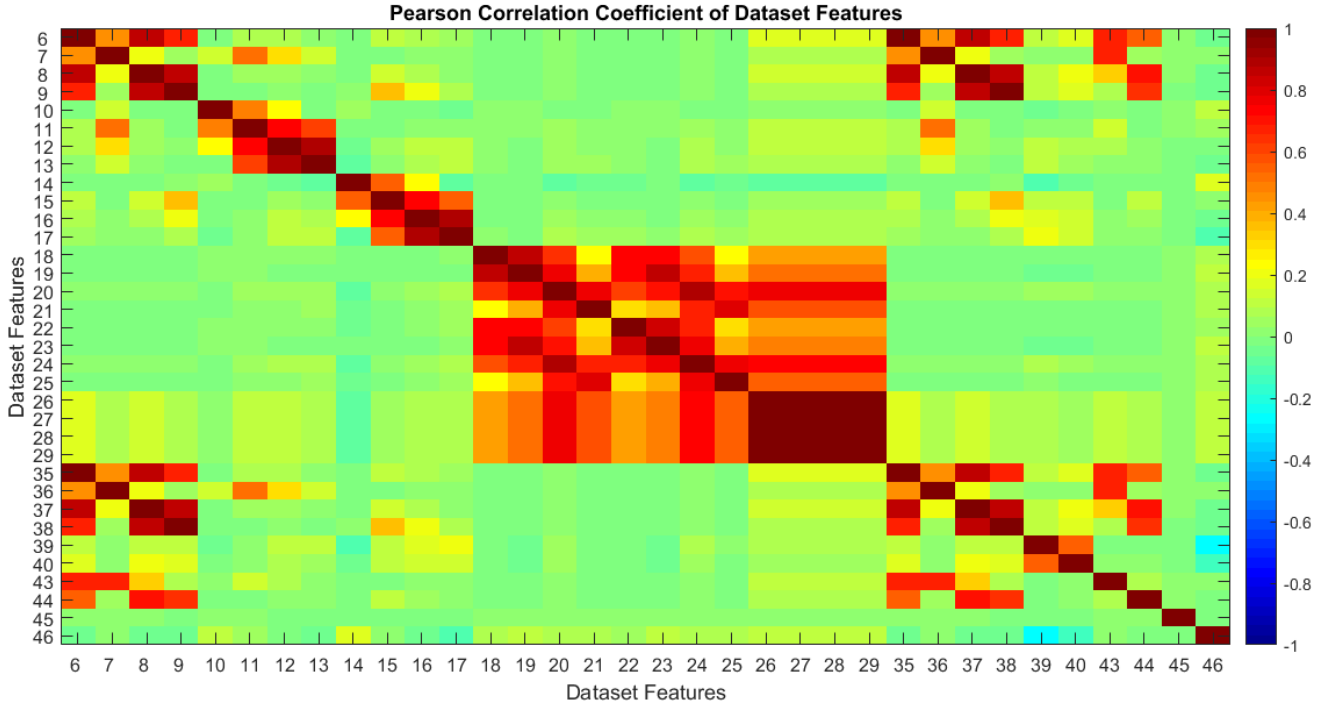


Fig. 1: Matrix of Pearson correlation coefficients of all features present in the dataset.

Spyware: Malware used to obtain information, which can record the online activities of the infected user, such as access credentials, credit cards and other financial information.

Ransomware: By blocking and encrypting the information on the infected machine, the attacker demands a payment, often in cryptocurrencies, so that the user can access his data again.

Virus: Malware that hides in benign files. It can modify other computer programs and infect them with its own code, spreading while damaging the system, being able to delete and corrupt files.

Trojan: Disguised as a legitimate program, the Trojan's goal is to gain unauthorized access to the system. They are used as a gateway to other malware, such as spyware and ransomware, when creating backdoors.

Worms: They are intended to spread via the network interface and may corrupt files. The more machines are infected, the faster the Worms proliferate. They can demand bandwidths that overwhelm networks, in addition to creating backdoors that leave computers vulnerable to other attacks.

Cross Site Scripting: These are malicious code injection attacks on benign websites in order to steal information and credentials from users. Generally, failures in validating user input data and the response from the Web server are the causes of these attacks. This type of attack exploits the trust that a user has in a website, causing the user to run scripts created by the attacker believing they are native to the benign website.

5.2 Transport Layer

The flows using TCP and UDP protocols were also analyzed. Among the observed attacks, there are the scanning of ports, hosts and version of services in use and flooding attacks.

Vulnerability Scanning - Ports and Service Version: Scanning tools can be executed by legitimate users, but results in the extraction of information in search of vulnerabilities to carry out an attack in the future. When scanning doors, the aim is to identify the status of the doors, which may be closed, listening or open. Service version scanning has the same purpose, checking if the victim has any outdated services with known vulnerabilities.

Flooding attack: It is a denial of service attack that causes a high volume of traffic on the victim's system, sending several UDP (User Datagram Protocol) or ICMP (Internet Control Message Protocol) packets. The attack congests the bandwidth of the attacked machine, which tries to process requests from the attacking machine, without being able to process legitimate requests. This type of attack can slow the system down, being able to bring it down completely in worse situations.

5.3 Network Layer

In summary, attacks such as IP spoofing and man-in-the-middle attacks were observed, usually created from the poisoning of ARP tables and vulnerabilities caused by the execution of malicious programs.

IP Spoofing: It is an attack aimed at masking IP packets using spoofed sender addresses, being used as a starting point for other attacks such as denial of service and data exfiltration. An attacker modifies the source address in the packet header, making the victim believe that the packet is from a legitimate source.

Man-in-the-middle attack: It is a group of attacks that is based on a malicious intermediary between the interaction of two parties. This intermediate can read, block and change the information and data exchanged between two users, hosts or servers that are communicating.

The labeling of the dataset flows as legitimate or malicious, necessary for the creation of models in supervised machine learning algorithms, was done through IDS Suricata. The dataset was also balanced to avoid bias during the model training and test phases, therefore being composed of equal parts of legitimate and malicious traffic.

6 The proposed architecture

The proposed system has two operation modes: online and offline. The online mode performs classification in real time, whilst the offline mode allows to observe the performance of multiple classifiers for a given dataset, making the resulting metrics available in the visualization module.

The proposed architecture, shown in Figure 2, is modular and consists of three main modules: data collection, processing and visualization.

The data collection module captures and abstracts network traffic flows. It also stores the datasets used in offline processing. The capture process reflects network traffic through the libpcap library. Then, the flowtbag tool abstracts the sequence of packets in the flows and their 40 features, including the flows length and the total number of packages for each flow. We use the five fields of the TCP/IP packet header, source IP address, destination IP address, source port, destination port, and protocol to abstract packets into flows. A channel on the Apache Kafka platform, which acts as a data buffer, receives the streams of data. We use the Hadoop Distributed File System (HDFS), a distributed database, to store the datasets used to train and test the classification models.

The processing module carries out the process of classifying these flows. The processing module is implemented in an Apache Spark cluster. This platform presents advantages to the development of the system, as it has libraries aimed at the implementation of machine learning algorithms and the fast processing of data in real-time, using the micro-batch method with the structured streaming engine. The training module extracts the classification model using a dataset labeled from HDFS. In the online mode of operation, packages are collected and added to an Apache Kafka channel, and flows are then classified as legitimate or malicious by the classification model obtained previously. For execution in the offline mode, the classification module runs tests on various algorithms and datasets, obtaining performance metrics for each combination. The results of the classification, both online and offline, are then sent to an Elasticsearch server using the Apache Spark integration library.

The visualization module allows the network administrator to visualize the classifications history and the current state of the network, as well as the results of tested algorithms. We implement the visualization module using the Elasticsearch[1] and Kibana[2] software, both developed by Elastic. Elasticsearch implements a distributed and efficient search server, based on JSON documents. It receives and stores the data as the processing module sends it after the classification process is finished. Kibana is responsible for providing a user interface through dashboards, displaying to the network administrator the data received by Elasticsearch in real-time for both execution modes. It also allows consultation by historical data, using the search server features Elasticsearch.

7 Performance Analysis

A cluster of four computers, one master and three slaves, using Ubuntu 19.04 operating system, composes the performance analysis environment. The master is a biprocessed Xeon X5570 with 4 cores and 96 GB of DDR3 RAM, and the slaves are biprocessed Xeon E5-2650 with 8 cores and 32 GB of DDR3 RAM.

The experiments uses accuracy, precision, sensitivity, F1-score and False Negative Rate (FNR) to evaluate the algorithms performance. Accuracy refers to the closeness of a measured value to a known value and precision refers to the closeness of two or more measurements to each other. Therefore, accuracy is given

[1] https://github.com/elastic/elasticsearch, accessed in April 2021.

[2] https://github.com/elastic/kibana, accessed in April 2021.

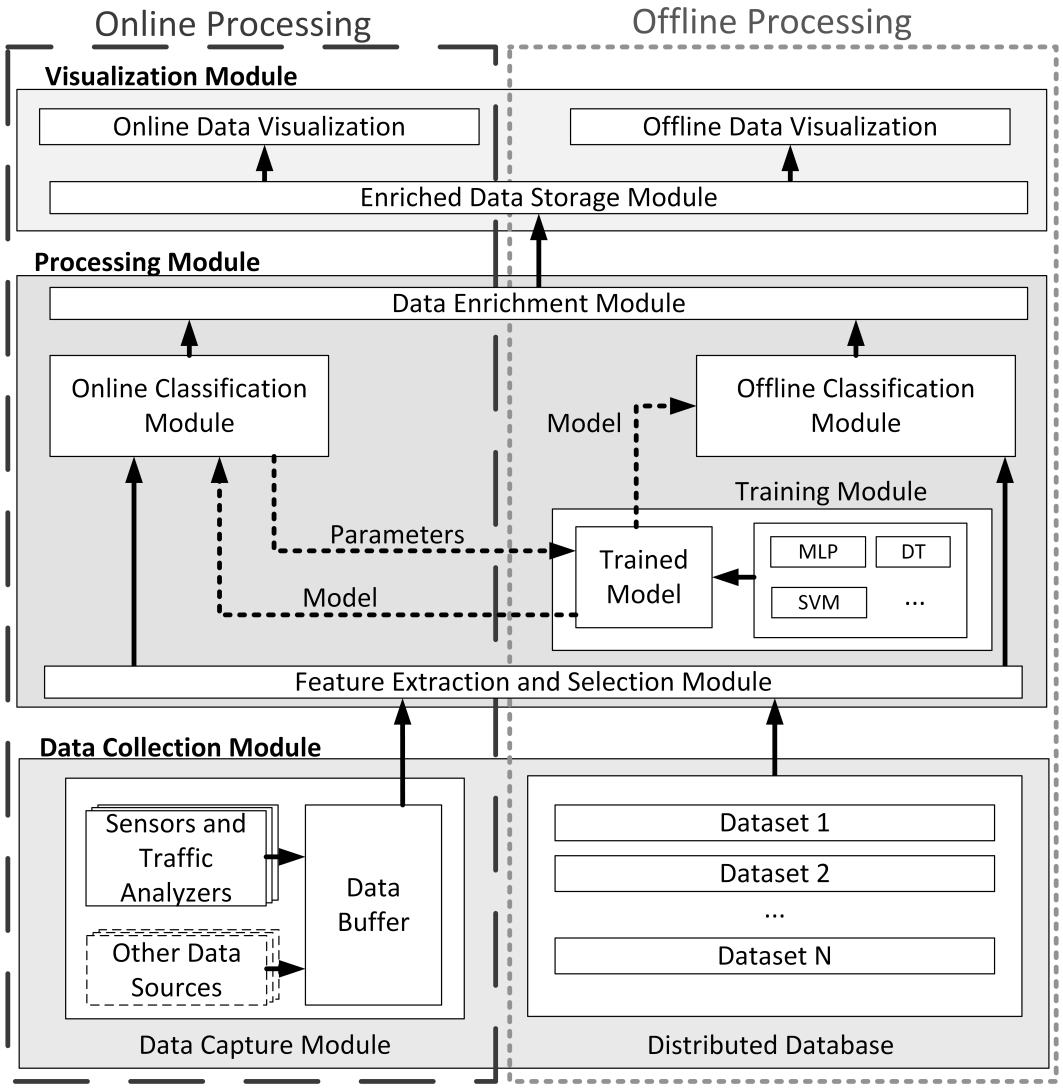


Fig. 2: TeMIA-NT modular architecture at online and offline modes.

by the number of flows correctly classified divided by total number of flows. High accuracy means that positively rated flows are less likely to be negative. Precision calculates the ratio of positive flows correctly classified among all flows classified as positive. Sensitivity calculates the proportion of all positive flows correctly classified among the actual positive flows. A high sensitivity means that most of the real positive flows have been classified correctly. The F1-score is the harmonic mean of precision and sensitivity, being a metric that takes into account false negatives and positives. The False Negative Rate detects the amount of false negatives, being equal to 1 minus the recall. Their equations are

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

$$Precision = \frac{TP}{TP + FP} \quad (4)$$

$$Recall = \frac{TP}{TP + FN} \quad (5)$$

$$F1\ score = \frac{2}{\frac{1}{precision} + \frac{1}{recall}} \quad (6)$$

$$False\ Negative\ Rate = \frac{FN}{TP + FN} \quad (7)$$

where TP = True Positives; TN = True Negatives, FP= False Positives, and FN = False Negatives.

7.1 Hyperparameter Tuning

In order to increase the model performance, hyperparameter tuning was applied using Apache Spark's implementation of the grid method. This method requires that all values for a given hyperparameter are set previously; the method then tests the performance of every possible combination of the given hyperparameter values, and returns the model that offers the best performance in a predetermined performance metric. The values tested for each hyperparameter are presented on Table 2, with Spark's default values highlighted. The target metrics chosen were the precision, to minimize the false positive rate, and F1-Score, to obtain a balance between both false positives and false negatives. The impact observed due to hyperparameter tuning varied depending on the algorithm, but in all cases the resulting model after tuning was equal to or better than the model without optimization.

Table 2: Algorithms tested and their hyperparameter values used on grid optimization.

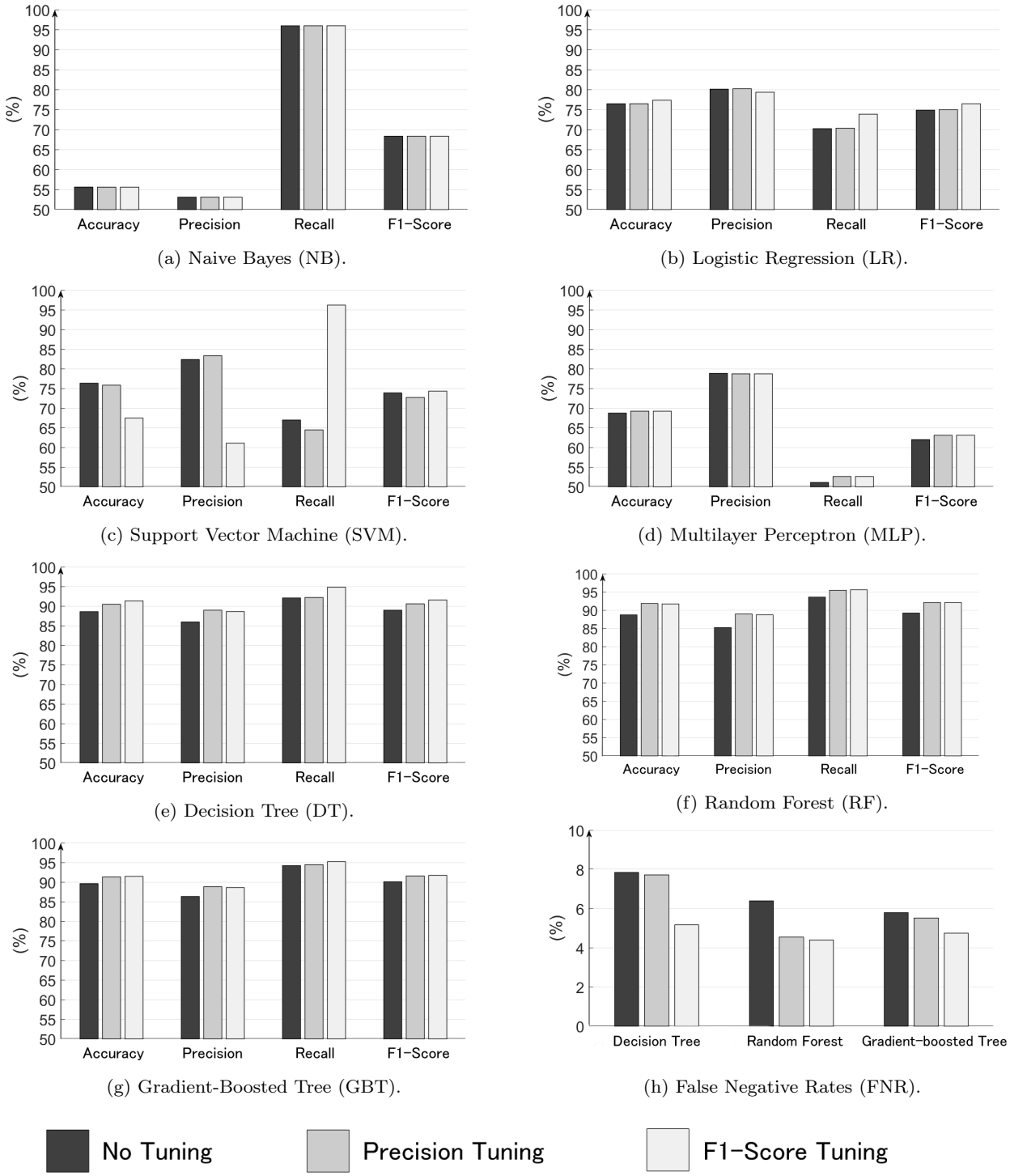
Naïve Bayes	
Smoothing	[0.0,0.25,0.5,0.75,**1.0**,2.5,5.0]
Logistic Regression	
ElasticNet Param	[**0.0**,0.25,0.5,0.75,1.0]
Max Iterations	[5,10,20,50,**100**,200]
Regularization Param	[**0.0**,0.01,0.1,0.3]
Tolerance	[**10^{-6}**,10^{-5},10^{-4},10^{-3}]
Support Vector Machine	
Fit Intercept	[**True**, False]
Max Iterations	[5,10,20,50,**100**,200]
Regularization Param	[**0.0**,0.01,0.1,0.3]
Standardization	[**True**, False]
Tolerance	[**10^{-6}**,10^{-5},10^{-4},10^{-3}]
Multilayer Perceptron	
Max Iterations	[5,10,20,50,**100**,200]
Step Size	[0.001,0.01,**0.03**,0.1]
Tolerance	[**10^{-6}**,10^{-5},10^{-4},10^{-3}]
Decision Tree	
Impurity	[**"Gini"**, "Entropy"]
Max Depth	[3,**5**,6,9,12,15,18,21,24,27,30]
Min Info Gain	[**0.0**,0.1,0.2,0.3]
Min Instances per Node	[**1**,2,5,10,20,40]
Random Forest	
Max Depth	[3,**5**,6,9,12,15,18,21,24,27,30]
Number of Trees	[**20**, 50, 100, 200, 300, 400, 500]
Subsampling Rate	[0.5,0.75,**1.0**]
Gradient-boosted Tree	
Max Depth	[3,**5**,6,9,12,15,18,21,24,27,30]
Step Size	[0.05,**0.1**,0.2]
Subsampling Rate	[0.5,0.75,**1.0**]

Figure 3a presents the results of hyperparametric optimization for naive Bayes. The optimal results for all tested cases, both with and without optimization, were the same. This occurred because the default value offered by the platform for the adjusted hyperparameter, smoothing, already offers the best results for the dataset used. As can be seen in Figure 3a, the model has high sensitivity, but the precision results are close to 50%. This indicates that the model has a high rate of false positives, which for an application aimed at detecting network threats is catastrophic, as it results in almost half of the normal flows being erroneously classified as malicious traffic. This demonstrates the importance of analyzing multiple metrics in building a model, as well as why F1-Score is considered a more complete metric than just accuracy or precision.

Figure 3b presents the performance results obtained for logistic regression. In the case of optimization for precision, the optimal hyperparameters follow the standard values, with the only difference being a convergence tolerance of 10^{-3}. By acting as a stopping criterion, a lower tolerance of convergence potentially results in a model with less overfitting, slightly improving the classification performance. Optimization for F1-Score, on the other hand, offered a more significant result, using 10 maximum iterations, a regularization parameter 0.3, and 10^{-3} convergence tolerance instead of the default values of 100, 0.0 and 10^{-6}, respectively. Using a smaller number of iterations offers a less accurate model, but the greater sensitivity and lesser overfitting resulting from changes in the other two hyperparameters compensate for this loss in the calculation of the F1-Score.

Figure 3c shows the performance results obtained for SVM. Similar to logistic regression, the adjusted hyperparameters seek to define a better stop condition for the algorithm. In optimizing for precision, the optimal model reduced the maximum number of iterations to 50, in addition to also reducing the convergence tolerance to 10^{-3}, both modifications to reduce overfitting. In the optimization for F1-Score, the maximum number of iterations and the convergence tolerance has also been reduced to 10 and 10^{-3}, respectively. Regularization was also applied, with the regularization parameter set to 0.3.

Figure 3d presents the results obtained after optimizing the multilayer perceptron. Similarly to the naive Bayes case, optimizations targeting both precision and F1-Score as the target metric offered the same results. In both cases, the optimal result was achieved by increasing the maximum amount of iterations from 100 to 200, as well as reducing the step size from 0.03 to 0.01. Both results demonstrate how a smaller step size



(a) Naive Bayes (NB).

(b) Logistic Regression (LR).

(c) Support Vector Machine (SVM).

(d) Multilayer Perceptron (MLP).

(e) Decision Tree (DT).

(f) Random Forest (RF).

(g) Gradient-Boosted Tree (GBT).

(h) False Negative Rates (FNR).

Fig. 3: Performance of the metrics of accuracy, precision, sensitivity, F1-score and FNR for the evaluated classifiers, for cases without optimization and with optimization of precision and F1-Score as target metrics. The FNR is only calculated for the classifiers with satisfactory performance in the precision and F1-Score metrics.

offers models with better performance without necessarily resulting in overfitting, since the convergence tolerance remained 10^{-6} even with the larger number of iterations.

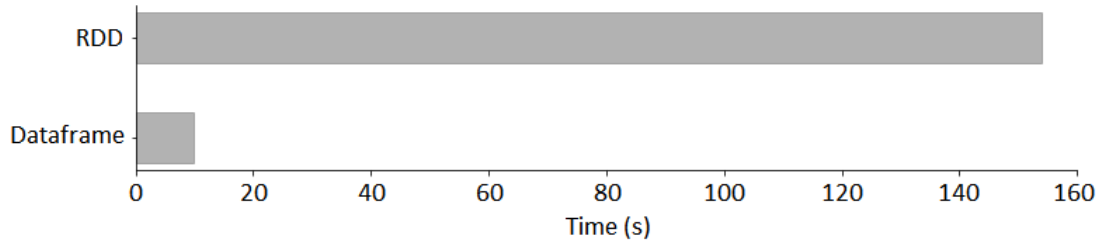


Fig. 4: Impact of the data structure on the training time of the decision tree.

Figure 3e presents the performance results for the decision tree. When optimizing for precision, the only hyperparameter with a value other than the default was the maximum depth, which went from 5 to 30. In F1-Score optimization, the maximum depth went from 5 to 18, and the method used to calculate the impurity was entropy, instead of the Gini method. Both results demonstrate the importance of the maximum depth hyperparameter, with the tree provided by default by the platform not offering a satisfactory performance. The model optimized for F1-Score also offered the greatest gain in FNR when compared to cases without optimization, as shown in Figure 3h; this model presented a percentage gain of 34.18%.

Figure 3f presents the optimization results obtained for the random forest algorithm. For optimization with precision as a target metric, the optimal values of hyperparameters were: maximum depth 30, number of trees 50 and subsampling rate of 0.75. For the optimization aiming at F1-Score, these values were: maximum depth 24 and number of trees 300. In both cases the values adjusted for maximum depth and number of trees are higher than the standard values offered by Apache Spark. Figure 3h demonstrates how this algorithm offers the lowest FNR after hyperparametric optimization; this reduction was accompanied by the second highest percentage gain: 31.46% when compared to the case without optimization.

Figure 3g shows the results of the gradient driven tree. The hyperparameters that offered the best results were, for optimization with precision as a target metric: maximum depth 18, step size equal to 0.2 and subsampling rate equal to 0.75. For the optimization aiming at F1-Score, the only hyperparameter with final value different from the standard values was the maximum depth, which went from 5 to 15. This reinforces the importance of the depth of the tree in the performance of this family of algorithms, being the only hyperparameter modified to result in the model with the highest performance in the selected metric in all tree-based algorithms. While the random forest shows the best performance after optimization, the gradient-boosted tree shows the best performance using the standard values of hyperparameters, considering both FNR and F1-Score.

7.2 Results and Analysis

The model convergence and training time plus the processing speed must be considered in the context of real-time analysis. To verify the impact of the data structure used during model training, we compared the model training time of Dataframe-based TeMIA-NT with the RDD-based CATRACA, and IDS previously proposed by our research group. Figure 4 shows that the DataFrame data structure several performance optimizations have a significant impact on the latency. Training the model with DataFrame is ten times faster than the same operation made with RDD.

We split the dataset into 70% for the training set and the other 30% for the test set to obtain the models in the offline processing mode. Also, we used K-fold cross-validation, with $k = 10$, to guarantee the model's generalizability. Finally, we use the grid search method to tune the hyperparameters in each algorithm. Figure 5 shows the results of each algorithm with weighted F1-Score set as the target classification metric; based on Figure 5, random forest, decision tree, and gradient-boosted tree models offer the best performance.

The online mode of operation uses the model with the highest processing capacity and good accuracy. Table 3 shows that the decision tree algorithm presented the maximum flow volume rate of 50 GB/s. The random forest classification model has a lower performance due to the need to process multiple trees, and it is necessary to obtain the result for all trees to achieve the final classification result.

As the decision tree model presents the best results both in accuracy and in classification capacity, this is the model used by default in the execution of the proposed system. However, other models can also be used according to the user's needs.

The last test observed the impact of parallelism on the system's performance, observing how variations in

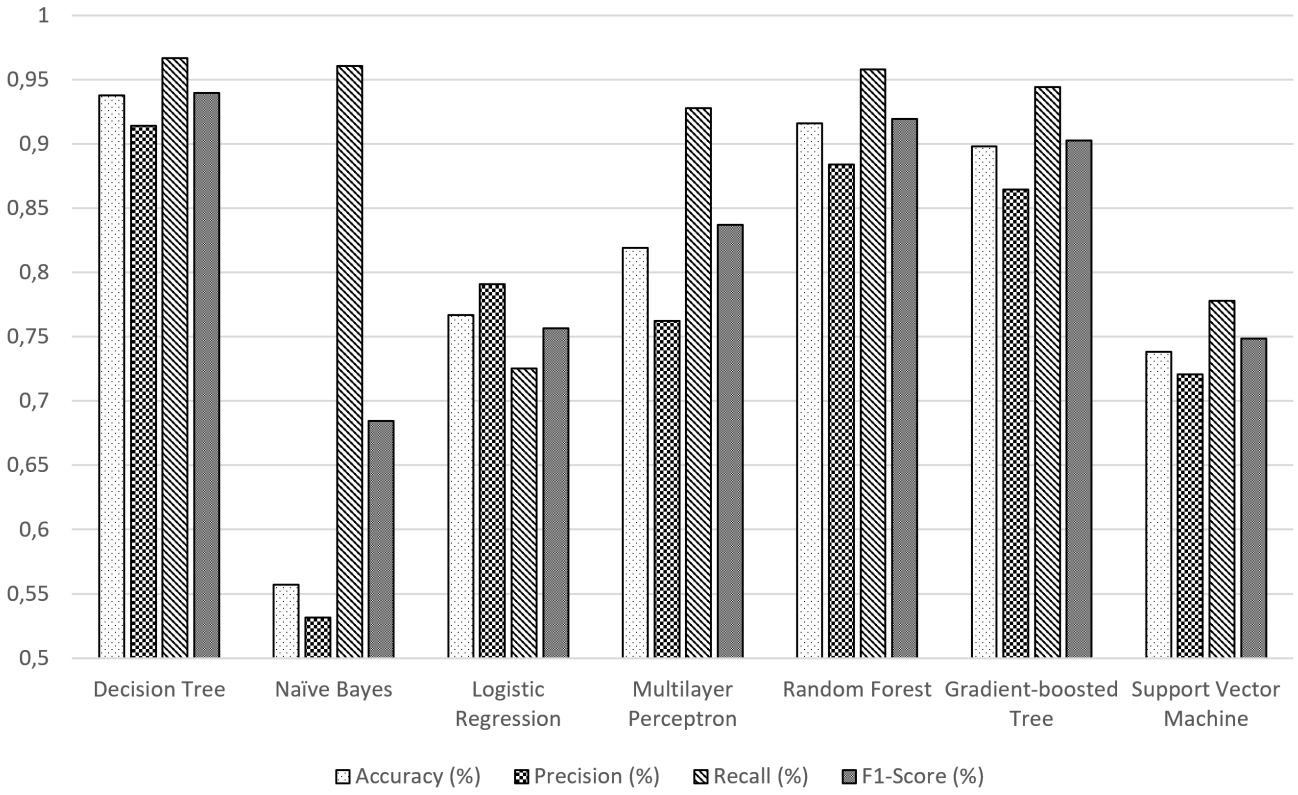


Fig. 5: Comparison of the evaluation metrics for the seven classifiers.

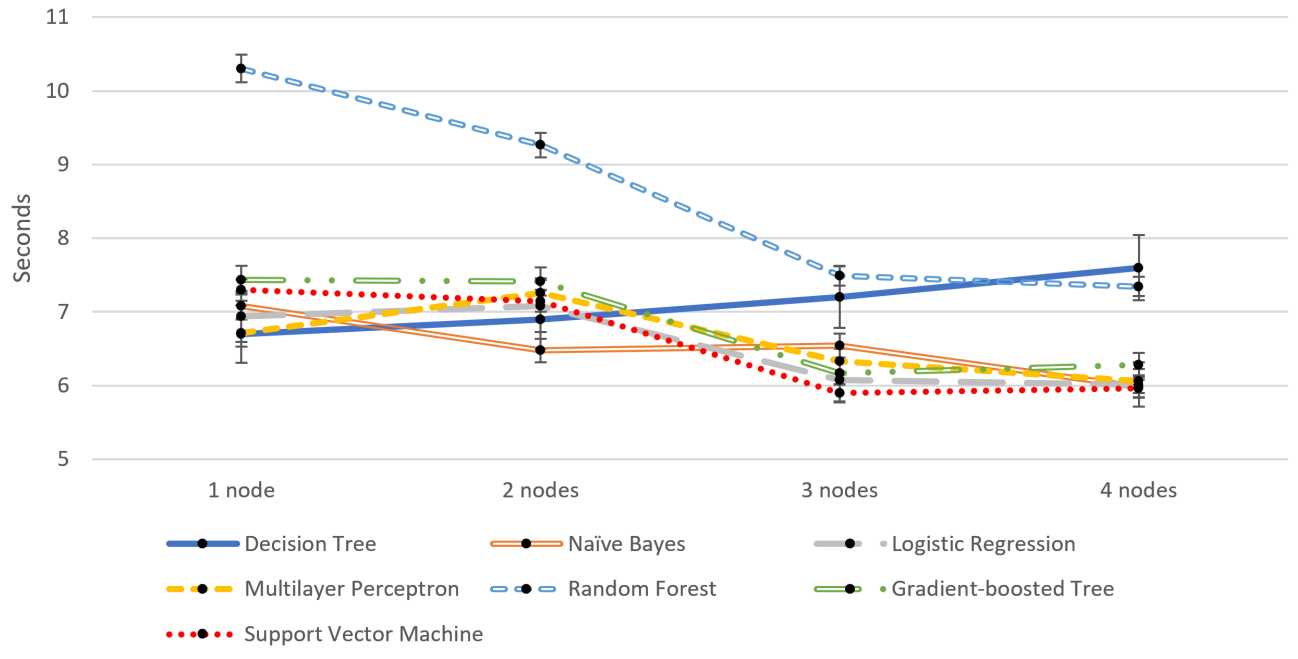


Fig. 6: Time necessary for classification based on number of processing nodes.

Table 3: Models processing efficiency with the best results in terms of the number and volume of classified flows per second.

	Flows/s	GB/s
Random Forest	586.563,32	21,95
Decision Tree	1.330.732,59	49,80
Gradient Boosted Tree	1.206.962,94	45,17

the number of processing nodes affect the results. The processing time required to classify 10 million network flows was measured while varying the number of processing nodes between 1 and 4 in the Spark environment; the results can be seen in Figure 6.

As can be seen from Figure 6, increasing the number of processing nodes reduces the total time required to process flows for most algorithms. This impact is more significant in the case of the Random Forest, as this algorithm works by creating multiple models of trees that can be executed in parallel during the classification process. The results of the Decision Tree algorithm are negatively affected by the increase in the number of nodes.

8 Conclusion

This article presents the TeMIA-NT system[3], developed to monitor traffic using parallel flow processing. TeMIA-NT presents two modes of operation: online and offline. The online mode of operation allows the network manager to monitor and detect network security threats in real-time. The offline mode of operation allows the performance evaluation of multiple classification models obtained from different algorithms and datasets. TeMIA-NT also allows the selection from seven machine learning algorithms when obtaining classification models. The detection of threats in real-time with low latency is achieved thanks to the dataframe data structure and the continuous processing engine of the structured streaming library.

The obtained results from a dataset based on legitimate traffic demonstrate the high processing capacity in flows per second. The performance of each implemented machine learning algorithm is also observed, with the decision tree and random forest models presenting high values in metrics such as accuracy and f1-score. These two algorithms presented a FNR performance increase of more than 30% when compared to the cases without optimization. This reduction was obtained while maintaining good precision and F1-Score rates, indicating that the models did not result in an excessive amount of false positives during the classification. It was also shown how most algorithms scale with an increasing number of nodes on an Apache Spark cluster.

[3] The code, documentation, and license are available at: https://www.gta.ufrj.br/TeMIA-NT/.

9 Acknowledgment

This work was financed by CNPq, CAPES, FAPERJ, and FAPESP (2018/23292-0, 15/24485-9, 14/50937-1).

References

1. "Cybersecurity Market Report. Available at: *https://cybersecurityventures.com/*. Last access: 30 April 2021."
2. E. Bertino and N. Islam, "Botnets and internet of things security," *Computer*, vol. 50, no. 2, pp. 76–79, 2017.
3. A. Azmoodeh, A. Dehghantanha, and K.-K. R. Choo, "Big data and internet of things security and forensics: Challenges and opportunities," in *Handbook of Big Data and IoT Security*. Springer, 2019, pp. 1–4.
4. Symantec, "Internet Security Threat Report. Available at: *https://docs.broadcom.com/doc/istr-24-2019-en*. Last access: 30 April 2021," 2019.
5. R. A. A. Habeeb, F. Nasaruddin, A. Gani, I. A. T. Hashem, E. Ahmed, and M. Imran, "Real-time big data processing for anomaly detection: A survey," *International Journal of Information Management*, vol. 45, pp. 289–307, 2019.
6. Verizon Enterprise, "Data breach investigations report. Available at: *https://enterprise.verizon.com/resources/reports/2020-data-breach-investigations-report.pdf*. Last access: 30 April 2021," 2020.
7. M. A. Lopez, D. M. F. Mattos, O. C. M. B. Duarte, and G. Pujolle, "Toward a monitoring and threat detection system based on stream processing as a virtual network function for big data," *Concurrency and Computation: Practice and Experience*, vol. 31, no. 20, p. e5344, 2019.
8. M. Pelloso, A. Vergutz, A. Santos, and M. Nogueira, "A self-adaptable system for DDoS attack prediction based on the metastability theory," in *2018 IEEE Global Communications Conference (GLOBECOM)*, 2018, pp. 1–6.
9. E. Viegas, A. Santin, A. Bessani, and N. Neves, "Bigflow: Real-time and reliable anomaly-based intrusion detection for high-speed networks," *Future Generation Computer Systems*, vol. 93, pp. 473–485, 2019.
10. R. Campiolo, L. A. F. dos Santos, W. A. Monteverde, E. G. Suca, and D. M. Batista, "Uma arquitetura para detecção de ameaças cibernéticas baseada na análise de grandes volumes de dados," in *Anais do I Workshop de Segurança Cibernética em Dispositivos Conectados*. SBC, 2018.
11. A. G. P. Lobato, M. A. Lopez, I. J. Sanz, A. A. Cardenas, O. C. M. Duarte, and G. Pujolle, "An adaptive real-time architecture for zero-day threat detection," in *2018 IEEE international conference on communications (ICC)*. IEEE, 2018, pp. 1–6.
12. M. A. Lopez, D. M. F. Mattos, and O. C. M. B. Duarte, "An elastic intrusion detection system for software networks," *Annals of Telecommunications*, vol. 71, no. 11-12, pp. 595–605, 2016.

13. M. A. Lopez, D. M. F. Mattos, O. C. M. B. Duarte, and G. Pujolle, "A fast unsupervised preprocessing method for network monitoring," *Annals of Telecommunications*, vol. 74, no. 3-4, pp. 139–155, 2019.
14. Cisco Systems, "OpenSOC: The Open Security Operations Center. Available at: *https://opensoc.github.io/*. Last access: 30 April 2021," 2014.
15. Apache Software Foundation, "Apache Metron. *https://metron.apache.org/*. Last access: 30 April 2021," 2017.
16. M. Zaharia, M. Chowdhury, M. J. Franklin, S. Shenker, and I. Stoica, "Spark: Cluster computing with working sets," *HotCloud*, vol. 10, no. 10-10, p. 95, 2010.
17. T. Jirsik, M. Cermak, D. Tovarnak, and P. Celeda, "Toward Stream-Based IP Flow Analysis," *IEEE Communications Magazine*, vol. 55, no. 7, pp. 70–76, 2017.
18. M. Zaharia, R. S. Xin, P. Wendell, T. Das, M. Armbrust, A. Dave, X. Meng, J. Rosen, S. Venkataraman, M. J. Franklin *et al.*, "Apache Spark: a unified engine for big data processing," *Communications of the ACM*, vol. 59, no. 11, pp. 56–65, 2016.
19. R. Xin and J. Rosen, "Project Tungsten: Bringing Apache Spark Closer to Bare Metal. *Available at: https://databricks.com/blog/2015/04/28/project-tungsten-bringing-spark-closer-to-bare-metal.html*. Last access: 30 April 2021," 2015.
20. M. Armbrust, R. S. Xin, C. Lian, Y. Huai, D. Liu, J. K. Bradley, X. Meng, T. Kaftan, M. J. Franklin, A. Ghodsi *et al.*, "Spark SQL: Relational data processing in Spark," in *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*. ACM, 2015, pp. 1383–1394.
21. X. Meng, J. Bradley, B. Yavuz, E. Sparks, S. Venkataraman, D. Liu, J. Freeman, D. Tsai, M. Amde, S. Owen *et al.*, "Mllib: Machine learning in apache spark," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 1235–1241, 2016.
22. J. Friedman, T. Hastie, R. Tibshirani *et al.*, *The elements of statistical learning*. Springer series in statistics New York, 2001, vol. 1.
23. L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.
24. L. Breiman, "Bagging predictors," *Machine learning*, vol. 24, no. 2, pp. 123–140, 1996.
25. L. A. C. de Souza *et al.*, "DFedForest: Decentralized Federated Forest," in *2020 IEEE Blockchain*, 2020, pp. 90–97.
26. J. Chen, K. Li, Z. Tang, K. Bilal, S. Yu, C. Weng, and K. Li, "A parallel random forest algorithm for Big Data in a Spark cloud computing environment," *IEEE TPDS*, vol. 28, no. 4, pp. 919–933, 2016.
27. M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, "A detailed analysis of the KDD CUP 99 data set," in *2009 IEEE Symposium on CISDA*, July 2009, pp. 1–6.
28. M. A. Lopez, R. S. Silva, I. D. Alvarenga, G. A. F. Rebello, I. J. Sanz, A. G. Lobato, D. M. F. Mattos, O. C. Duarte, and G. Pujolle, "Collecting and characterizing a real broadband access network traffic dataset," in *2017 1st Cyber Security in Networking Conference (CSNet)*, Oct 2017, pp. 1–8.
29. Daniel Arndt, "Flowtbag. *Available at: https://github.com/DanielArndt/flowtbag/wiki/ features*. Last access: 30 April 2021," 2011.