Université de Namur
Faculté des Sciences
Département d'informatique

Report : Flagging suspicious hosts from DNS traces

Luyckx Marco 496283
Bouhnine Ayoub 500048

Professor: ROCHET Florentin

Course: Data analysis for cybersecurity ICYBM201

Delivered in November 2023

Academic year 2023-2024

Contents

1 Introduction

This project aims to develop a classifier that is able to distinguish between 3 different kinds of traffic : human, bot and a combination of human and bot. Based on labeled datasets of bots behaviour and human behaviour, we aim to train a model to classify each host into one of the 3 category. This report will outline the methodologies employed, the challenges faced and the results obtained. Additionally, a critical discussion of the result will be given regarding the common pitfalls for learning-based IDS.

1.1 How to run the project ?

Initially, the professor gave us 2 python files, namely `train.py` and `eval.py`. We have restructured the project to enhance its coherence and scalability.

In the following commands, '<algo>' can be replaced by 'decision_tree', 'logistic_regression', 'neural_networks', 'random_forest' or 'knn'. It is not a mandatory argument for `train.py`, but if you do not specify it, the default algorithm will be 'logistic_regression'. We made it mandatory for `main.py` because this script is not required by the project statement.

The different commands that you can use are :

Training the model

Listing 1: Command for training the model

```
python3 train.py \
--webclients ../training_datasets/tcpdumps/webclients_tcpdump.txt \
--bots ../training_datasets/tcpdumps/bots_tcpdump.txt \
--algo <algo> \
--output ../trained_models/<algo>/trained_model_<algo>.pkl
```

where:

- `--webclients` specifies the path to a file containing data related to webclients;
- `--bots` specifies the path to a file containing data related to bots;
- `--algo` specifies the algorithm to use for the training process;
- `--output` defines the path where the trained model will be saved after the training process is completed. It needs to be in the Pickle format (`.pkl`).

Evaluating the model

Listing 2: Command for evaluating the model

```
python3 eval.py \
--trained_model ../trained_models/<algo>/trained_model_<algo>.pkl \
--dataset ../evaluation_datasets/tcpdumps/eval1_tcpdump.txt \
--output ../suspicious_hosts/suspicious_hosts.txt
```

where:

- `--trained_model` specifies the path to a pre-trained model which is in the `.pkl` format;
- `--dataset` specifies the path to the dataset that will be used for the evaluation;
- `--output` defines the path where the result of the evaluation will be saved. This file will contain a list of hosts considered suspicious by the trained model.

Training and evaluating simultaneously

Listing 3: Command for training and evaluating the model at the same time

```
python3 main.py \
--webclients ../training_datasets/tcpdumps/webclients_tcpdump.txt \
--bots ../training_datasets/tcpdumps/bots_tcpdump.txt \
--algo <algo> \
--trained_model ../trained_models/<algo>/trained_model_<algo>.pkl \
--dataset ../evaluation_datasets/tcpdumps/eval1_tcpdump.txt \
--output ../suspicious_hosts/suspicious_hosts.txt
```

The parameters for this command assume the same function as from the training and evaluation of the model.

This subsection only provides a concise guide. For more information and detailed instructions, refer to the main `README.md` file in the project's root directory.

2 Explanations on how we approached the problem

For everyone to start with the same data, the professor provided us with a set of datasets. We received two distinct training datasets : one containing traffic exclusively generated by bots, named `bots_tcpdump.txt`, and another containing traffic from human webclients, named as `webclients_tcpdump.txt`.

Also, we were given two evaluation datasets, each accompanied by a corresponding botlist. The first evaluation dataset, `eval1_tcpdump.txt`, consists of two distinct groups : bots and humans. These groups are guaranteed to be separate sets of hosts. In contrast, the second evaluation dataset, `eval2_tcpdump.txt`, is more complex, where certain hosts emit traffic from both humans and bots.

It's important to note that our machine learning models are exclusively **trained** using the **training datasets** and **evaluated** using the **evaluation datasets**. **There is no mixing of data between the two steps.**

2.1 Our workflow

1. **Understanding and exploring the data**
   - **Exploratory data analysis** : We explored the patterns in the datasets to understand them and identify differences between them. Some of the tasks included analyzing the average query length of the requests and responses, the average number of dots in a domain and some temporal patterns, etc...

2. **Data pre-processing**
    - **Feature engineering** : Based on what we found on the exploration data phase, we created new features that we believed would be relevant for classification. See section features selection for more details.
    - **Data cleaning** : We handled missing values, for example in cases where an ID is associated only with a response, indicating no corresponding request. In such cases, we removed the data. Additionally, we addressed situations where there were no responses after a request, considering the response as empty and filling it with zeros or 'None' as appropriate.
3. **Classifier selection**
    - **Baseline model** : We started with a simple classifier, such as logistic regression, to establish a baseline and understand the model's behavior and expected results.
    - **Other classifiers** : Then, we explored other classifiers, including decision trees, random forests, neural networks and KNN to compare their performance.
4. **Training and validation**
    - **Model training** : We trained the selected models using the training dataset.
    - **Model evaluation** : To validate the efficacity of our model, we performed a validation process using a dedicated testing set. This evaluation includes the computation of different metrics to measure various aspects of the performance. Among these metrics, we used the detection rate, the false alarm rate, the false positive rate and the true negative rate. Additionally, we used different statistical measures such as accuracy, performance, recall and F1-score. These metrics will provide an overview of model effectiveness from multiple perspectives, therefore, ensuring a robust assessment.
5. **Diagrams**
    - **Create diagrams** : To compare different metrics for each combination of features and classifiers, we generated different diagrams. These visualizations will help in the interpretation and comparison of results.

2.2 Format of DNS queries

The first challenge we encountered involved understanding the format of the provided data, enabling us to identify and extract relevant features necessary for our machine learning models.

We started by examining the structure of a typical DNS query from the tcpdumps that were given. Consider the following line :

```
13:22:44.546969 IP unamur021.55771 > one.one.one.one.domain:  18712+ A? kumparan.com.
(30)
```

Breaking this down, we find :

1. **13:22:44.546969** : This is the timestamp at which the DNS query was logged, indicating the exact date the event occurred.
2. **IP** : Specifies that this log entry is concerning an IP packet.
3. **unamur021.55771** : This is the source of the DNS request. `unamur021` is the hostname of the device making the request and `55771` is the source port number from which the request is coming. Since port numbers are chosen at random, they should **not** be considered as features.
4. **>** : Indicates the direction of the traffic.
5. **one.one.one.one.domain** : This is the destination of the DNS request.
6. **18712+** : This is the query ID for this DNS request. The '+' indicates that this is a recursive query.
7. **A?** : This indicates the type of DNS record being requested.
8. **kumparan.com.** : This is the domain name for which the A record is being requested.
9. **(30)** : This is the length of the DNS request packet in bytes.

When it comes to the **response of a DNS query**, the format remains mostly identical. However, there are some key differences :

- **X/Y/Z** : This is a breakdown of the answer sections in the DNS response (in the code, this will be referred to as `counts`) :
  1. **X** : indicates there are X answers.
  2. **Y** : indicates there are Y authoritative nameservers.
  3. **Z** : indicates there are Z additional records.
- **A 104.18.130.231, A 104.18.129.231** : These are the answers to the A record query for the earlier example query `kumparan.com` and it means that the domain has at least two IPv4 addresses associated with it : `104.18.130.231` and `104.18.129.231`.

Understanding the format of these queries is important for the future feature selection. We must identify features that are relevant and impactful for distinguishing between human and bot traffic.

2.3 Our assumptions and considerations

To effectively address the problem, we made certain assumptions to guide our approach. Some of these will be explain in details later during the pitfalls :

- As stated in the project statement, we exclusively consider the `1.1.1.1` DNS resolver.
- Our analysis is confined to DNS captures and our approach is specifically tailored for DNS data.
- Some responses may appear without a preceding request, which we consider as unusual behavior.

- Label inaccuracy (P2) : We assume that the datasets have been correctly labeled and that the labels are reliable for training our models.
- **Our classifiers categorize both bots and instances of mixed bot and human behavior as 'bot'**. We believe that an infected computer, regardless of whether it is used by a human, should be classified as a 'bot' but with a certain threshold. Thus, non-infected instances are labeled as 'human'. For a visual representation, refer to the Figure 1. There is a dedicated section later in the report where we will give more details about this threshold.
- For our purposes, a **'session'** encompasses the entire trace. When we introduce new features based on aggregated data , a **'user-session'**, (it does not matter whether the user is a bot or human) spans from the initial request sent to the final response received, regardless of whether the user is a bot or a human.
- We observed that in the provided datasets, sometimes there were flags from TCP connections. However, we decided that these flags would not be taken into account during our analysis and parsing, as the focus of this project is to flag suspicious hosts based on **DNS traces**.

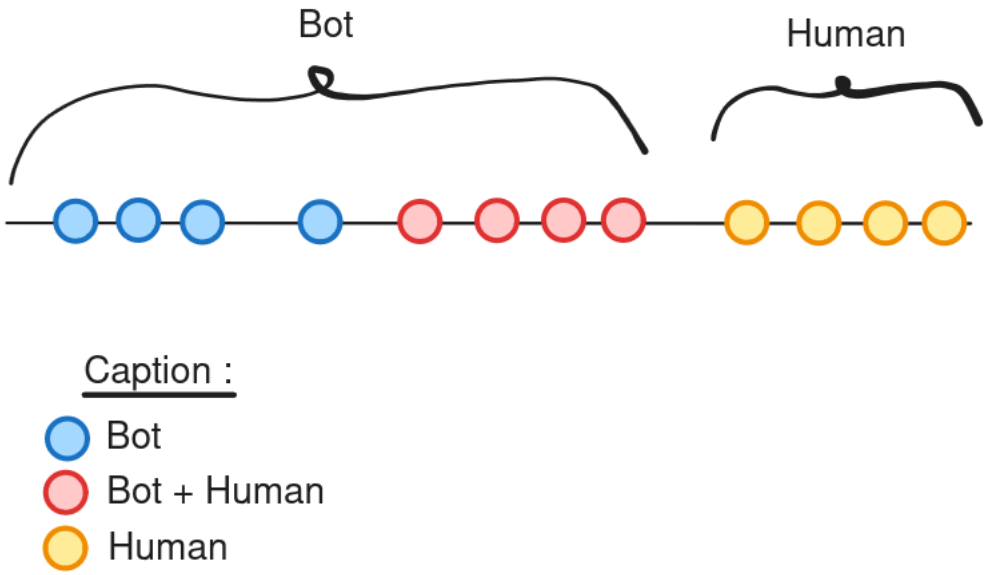

Figure 1: Simplified logistic regression for classification

2.4 Features selection

The development of an effective IDS based on ML requires careful consideration of the features that will be used for classification. Using only the raw features present in the datasets would not be optimal, as detecting bot **behavior** requires more than just examining **individual** queries (either requests or responses) in isolation.

Thus, to enhance the efficiency of our model, we opted to aggregate several raw features

extracted during the data pre-processing phase. These aggregated features provide a more accurate representation of host behavior.

To identify the distinguishing features, we examined the two training datasets by hand, aiming to find differences between human and bot traffic patterns. We found that:

- Human traffic tends to occur in bursts, concentrating activity within a few minutes and often on similar websites.
- Bot traffic, in contrast, is persistent yet sporadic. They perform requests at regular intervals (e.g., every 10 seconds) across a diverse range of websites.

2.4.1 Aggregation of data and combination of features

In our initial step, we begin by combining **each request-response pairs** based on their **ID**. If there are different requests-response pairs that have the same ID (it could be the case since we have a lot of data inside the dataset and therefore there is a probability to have the same ID for two distinct pair of request-response), we simply **append to the end of the ID a number** and each time we see again the same ID we **increment** this appended number. Once this completed, we aggregate **each pairs** to create a **behaviour** profile for **each host**. Following this aggregation, we then identify and extract useful features that will be explained below.

We created 3 different categories of features :

- **Features related to numbers**
  - **average number of dots in a domain** : Calculating the average number of dots in a domain queried by a host. Bots often query domains with **a single dot** (e.g., kumparan.com, fast.com), while webclients query **more complex domains**, including those queried by browsers (e.g., location.services.mozilla.com, services.addons.mozilla.org). The average number of dots serves as a distinguishing feature between bots and human clients.
  - **number of requests in a session** : Determining the number of requests in a session for each host. We define a user-session as the time between the initial request sent and the final response received. Webclients average around **1062** requests per session, whereas bots average approximately **22**. This very significant difference highlights the more active traffic of webclients compared to bots.
  - **number of unique domains** : Counting the number of unique domains queried by each host. We found that the bots query each time different domains compared to webclients who query in the most case several time usual domains. On average, webclients have around **275** unique domains, while bots have approximately **22**. Interestingly, the number of unique domain is almost the same as the number of requests in a session for the bots. Indeed, it confirms the assumption where almost all bots query each time different domains.
  - **average counts** : Computing the average number of answers received in a response for each host. This feature reveals that webclients receive an average of 1 **answer**

per request, while bots receive approximately 2 answers per request. Bots tend to request more answers than webclients for a single query.

- **Features related to time**
  - **average time for a session** : Determining the average time for a **user-session**. As a reminder, we defined a user-session as a the timing spent between the initial request sent by the host in the trace and the last response received by the host in the trace. The session duration for bots is approximately **875.78 seconds**, while for webclients, it is around **1113.61 seconds**.
  - **average time between requests** : Calculating the average time between each request for each host. We thought that this feature would be great because bots and webclients do not send requests at the same frequency. We therefore took all the differences in time between each request. We found that webclients send request every **2** seconds in average whereas bots send every **40** seconds. We concluded that the traffic for a webclient is more active that the traffic of a bot.
  - **frequency of repeated requests in a short time frame** : Measuring the frequency of repeated requests for the same domain within a short time frame for each host. This feature helps identify if a host makes consecutive requests for the same domain within a short time. Bots query the same domain at a frequency of **0.0119** on average within a short time frame, while webclients query at a frequency of **0.9715**. It is coherent since we found that the majority of bots does not query the same domain twice since the number of request and unique domain is in average nearly equal as seen before.
- **Miscellaneous features**
  - **average of request length** : Calculating the average length of each request sent by each host. On average, request lengths for bots are approximately **30**, while webclients have an average request length of around **40**, which makes it a good distinguishing feature.
  - **average of response length** : Similar to request length. On average, response lengths for bots range between **60 and 70**, while response lengths for webclients **vary widely**.
  - **type of requests queried by hosts** : Categorizing the types of queries made by hosts (e.g., "A", "AAAA"). It can be a good feature to separate bot from webclients since, we noticed that in the training dataset, bots tend to query only "A" requests, while webclients exhibit more diverse query types.
  - **type of responses received by hosts** : Categorizing the types of responses received by hosts (e.g., "A", "AAAA", "NXDomain", "CNAME", "ServFail"). This feature highlights the different response types received by hosts and helps in distinguishing between bot and human behavior. Webclients receive a variety of response types, while bots receive limited types.

One important observation we've made and which we've been careful not to fall into, concerns the treatment of timestamps. It's very important to note that we must **not** overemphasize the importance of **individual timestamp values**. Otherwise, it would introduce a bias into our model because the model could prioritize specific timestamp values, thus affecting the model's ability to discern genuine patterns and features :

1. **Timestamps for bots tcpdump**
   - First timestamp of the trace : 13:22:44.546969
   - Last timestamp of the trace : 14:35:23.695938
2. **Timestamps for webclients tcpdump**
   - First timestamp of the trace : 11:44:12.890000
   - Last timestamp of the trace : 13:16:33.136055

Instead of focusing on individual timestamp values, we incorporated them as part of a broader context, such as a session or an aggregation between a request and a response. In some sense, we prioritized a **session-level** structure over a **query-level** one.

To prepare for our future machine learning algorithms, we will need some relevant combination of features. However, with a total of 11 features to consider, attempting an **exhaustive search** for the best combination would be testing **2,047 possibilities**. In practical terms, this computation would be extremely time-consuming, even on powerful commercial-grade computers.

To put this into perspective, on Marco's computer, a **single run takes approximately 25.5 seconds**. When multiplied by 2,047 (25.5 * 2,047), the total time required amounts to **52,198.5 seconds**, which translates to roughly 869,975 minutes or about **14.5 hours for a single algorithm**. Given that we have **5** algorithms to evaluate, this process would demand a cumulative time of **72.5 hours** just to compare and select the optimal feature set. Unfortunately, as students, we do not have that much resources.

Instead, we adopted a more efficient approach by grouping together features that **logically made sense to us**. This led us to identify 7 distinct sets of features that were most suitable for our project :

1. **Time-related features** :
   - **Average time for a session**
   - **Average time between requests**
   - **Frequency of repeated requests in a short-time frame**

   Time-related features are interesting for distinguishing between bot and human traffic because they capture the behavioral rhythm of interactions with a network, which often differs between bots and human activities. Therefore, by understanding the temporal patterns unique to bots and humans, our IDS can be more effective in flagging suspicious activity and protecting the network against automated threats.

2. **Numbers-related features** :
    - **Average number of dots in a domain**
    - **Number of requests in a session**
    - **Number of unique domains**
    - **Average counts**

    Numbers-related features are essential in understanding the structural and quantitative aspects of network interactions. They provide different measures regarding the kind of domains with which hosts interact, which can be really important for identifying the systematic behavior of bots and human users. The model trained on these features could, therefore, differentiate between bots and humans based on patterns like domain complexity, domain diversity and the nature of the DNS interactions.

3. **Miscellaneous features** :
    - **Average of request length**
    - **Average of response length**
    - **Type of requests queried by hosts**
    - **Type of responses received by hosts**

    Miscellaneous features provide both qualitative and quantitative data about the requests and responses. These features are interesting because they encapsulate both the content and context of network interactions, providing a deeper look into the communication patterns between hosts and the network. By understanding the typical lengths and types of requests and responses, as well as their variances, our IDS can more accurately characterize and classify the network traffic, leading to more effective identification of bot-related threats.

4. **High-level behavioral features**:
    - **Average number of dots in a domain** : Differentiates based on domain query patterns.
    - **Number of requests in a session** : Reflects the activity level of the host.
    - **Average time for a session** : Indicates the overall session duration and can reflect user engagement or bot operational patterns.
    - **Frequency of repeated requests in a short time frame** : Helps to distinguish bots that may repeatedly query the same domain in automation tasks.
    - **Type of requests queried by hosts** : May highlight the simplicity or complexity of queries, indicative of automated or human-driven behavior.

    This combination focuses on general patterns of host behavior, capturing both the complexity of their web activity and the intensity of their interactions with domains. Features like the average number of dots in a domain and the type of requests depicts

the complexity of user browsing patterns against the more straightforward patterns of bots. The number of requests in a session and the average session time provide a really good insight of user activity in a session, where more requests and longer times likely indicate human activity. The frequency of repeated requests within a short timeframe could give an idea of whether the same domain is repeated in a given timeframe, which highlight the difference of frequency between bots and humans.

5. **Domain interaction and traffic patterns features**:
   - **Number of unique domains** : Bots tend to query different domains, providing a clear behavioral distinction.
   - **Average counts** : Differentiates based on how many answers are typically requested, which could signify automated processes.
   - **Average time between requests** : Indicates the regularity or irregularity of traffic, which can be a revealing sign of automated querying.
   - **Average of request length** : This could serve to differentiate between the complexity of interactions.
   - **Type of responses received by hosts** : Varied type of responses might indicate a more human-like interaction with the network.

   This combination give some insight about the interaction between the host and the domains it contacts, alongside the traffic patterns. It captures the diversity and frequency of domain queries (number of unique domains and average time between requests) to distinguish between the broader exploration of domains from humans and bots. The average request length can indicate the level of detail in the queries. The type of responses indicates whether the host's behavior solicits a broad or limited set of responses.

6. **Response-based and session duration features**:
   - **Average of response length** : Suggests the complexity or nature of the information being exchanged.
   - **Average time for a session** : Prolonged sessions might be more human-like, whereas shorter could imply bot activities.
   - **Average time between requests** : Having the same intervals between 2 queries could imply automation whereas more varied timings could indicate human behavior.
   - **Frequency of repeated requests in a short time frame** : By looking at the datasets, we inferred that low frequency may indicate bot behavior.

   This combination of feature focuses on the informational content and timing of sessions to discern bots from humans. Average response length and average time for a session are used to measure the depth and duration of interactions. The average time between requests offers insights into the rhythm of traffic, where irregular intervals might suggest human browsing, and regular ones could indicate bot operations. And, as explained

before, the frequency of repeated requests within a short timeframe could help us in identifying the frequency between bots and humans within a timeframe.

7. **Detailed request and response patterns features**:
   - **Type of requests queried by hosts** : Different query types might be indicative of a diversified human activity.
   - **Type of responses received by hosts** : A limited range of responses might indicate bot activity.
   - **Average of request length** : Reflects the complexity of the queries made by the host.
   - **Average of response length** : Could indicate if a bot is programmed to carry out specific tasks that receive responses within a predetermined size range.

   This last combination is useful to capture the nuanced details of host interactions through the type of DNS requests and responses and their lengths. The diversity in the types of requests and responses provides a overview of a host's network behavior, distinguishing patterns of human users from those of bots. The average lengths of requests and responses add an additional layer of detail by providing insights about the length of queries.

For the future diagrams and comparisons of algorithms, we choose to stick with these combinations and also with a **baseline combination containing all the features** defined before.

3 Background knowledge

3.1 Slides from the course

Prior to the project, we had a whole lesson concerning IDS and about 2 critical aspects of IDS :

- **Effectiveness** : This dimension measures the accuracy with which an IDS **correctly** identifies intrusions within network traffic.
- **Security** : We discussed around the notion that high accuracy alone does not guarantee security. In particular, we emphasized the significance of **bayesian detection rates** and underscored the counter-intuitive nature of **the base-rate fallacy**. This fallacy suggests that even with a high overall accuracy, a system can still yield a considerable number of false positive alarms. Such a scenario is undesirable and we must be mindful of it.

The most important key point was :

`"To be effective an IDS must achieve a high level of accuracy, but to be truly secure, it must also attain a high bayesian detection rate."`

This will be extremely important during the 'discussion about our results' section.

3.2 Articles

In addition to the course materials, we had to read further research by examining 2 articles closely linked to the project :

1. **The base-rate fallacy and the difficulty of intrusion detection**[2] : This article introduces and mainly focuses on the base-rate fallacy, a non-intuitive concept in the field of probability that significantly impacts the effectiveness of IDS. The key objective for a secure IDS is not only to detect as many actual intrusions as possible but also to limit the number of false positives. The paper argues that the frequency of these false positives is a major limiting factor in an IDS's performance, with the base-rate fallacy playing a crucial role in this limitation. Essentially, the base-rate fallacy happens when someone is overlooking the base rate of the least frequent class in cases of class imbalance, therefore, it can distort the perceived effectiveness of a predictive model. In situations where the negative class is predominant, even a model with a low false-positive rate can produce a surprisingly large volume of false positives, undermining the model's apparent accuracy.

   We are going to develop in depth the base-rate fallacy and its implications in the next section.

2. **Dos and Don'ts of machine learning in computer security**[1] : This paper examines the application of machine learning in the field of computer security. The authors identify common pitfalls in the design, implementation and evaluation of learning-based security systems through a study of 30 papers from top-tier security conferences over the past decade. They find that these pitfalls are widespread and can lead to over-optimistic results, undermining the performance and practical deployment of ML in security tasks.

   The authors categorize 10 common pitfalls into different stages of the ML workflow :

   - **Data collection and labeling** :
     - **Sampling bias (P1)**
     - **Label inaccuracy (P2)**
   - **System design and learning** :
     - **Data snooping (P3)**
     - **Spurious correlations (P4)**
     - **Biased parameter selection (P5)**
   - **Performance evaluation** :
     - **Inappropriate baseline (P6)**
     - **Inappropriate performance measures (P7)**
     - **Base rate fallacy (P8)**
   - **Deployment and operation** :

– **Lab-only evaluation (P9)**

– **Inappropriate threat model (P10)**

The paper emphasizes the importance of recognizing and mitigating these pitfalls to avoid biased results and over-optimistic conclusions. It provides actionable recommendations for researchers to support the avoidance of these pitfalls where possible.

3.3 Pitfalls

As stated earlier, from the second article[1], we learnt a lot about pitfalls in machine learning. Understanding these pitfalls is really important for our project since we need to take them into account during our analysis and it will serve as a shield against making erroneous assumptions.

In the next sections, we will dive deeper into the specifics of these pitfalls, how they apply to our project and what we did to identify them.

3.4 Data collection and labelling

The first step for the design and development of a machine learning IDS is to obtain a representative dataset. From this stage, we need to consider the two following pitfalls : **sampling bias** and **label inaccuracy**.

3.4.1 Sampling bias

Sampling bias in the context of an IDS refers to a systematic error that occurs when the data used to train or test the IDS is not representative of the true population of network traffic or security events that the IDS will encounter in operations.

In our case, we used datasets provided by the professor, namely 'bots_tcpdump.txt' and 'webclients_tcpdump.txt'. The question we must ask ourselves is, "Do these datasets genuinely mirror real-world practices (which in our case are the dataset evaluations) ?" If there are not representative of the real-world distribution of human and bot DNS traffic, then our model might not perform well in practice.

Remarks :

- The limitation of the datasets provided lies in the fact that it has a limited number of DNS traces which do not really represent the true data distribution. Indeed, these traces are restricted to a specific range of hours (within a single day) and as a result, they do not adequately represent the whole network traffic. While efforts can be made to mitigate this bias, it is practically impossible to entirely represent the whole network behavior across all possible scenarios. Since this was given by the teacher, we cannot reduce this issue.
- Additionally, the bots identified within the DNS trace demonstrate a specific pattern of DNS queries. Thus, if a different type of botnet occurs in the network, the IDS might not successfully detect these bots, as they would likely generate a distinct pattern of

requests. As a result, the data would differ, potentially causing our classifier to miss the wider range of bot behaviors.

Very important remark : in the project statement, we are tasked with assessing our classifier on the **eval2** dataset, where hosts may behave like **bots at certain times** and **humans at others**. However, the training datasets we have been given only contain distinct instances of bot behavior and human behavior, without any overlap. It will have a significant impact on the performance of the model, more details will be given later.

3.4.2 Label inaccuracy

Label inaccuracy refers to the incorrectness of the labels assigned to data points in a dataset. Accurate labels are really important for classification machine learning algorithms since they rely on these labels to learn the relationship between the input data and the desired output. If it fails to do so, the overall performance of the model will be affected.

Given that the training datasets were provided by our professor, it is reasonable to assume that the ground-truth labels are accurate because if they are not, it would necessitate to relabel all the datasets which would be very challenging. Consequently, we are making the assumption that the professor has correctly labeled the datasets.

Remarks :

- If the labeled datasets were classified based on certain heuristics or assumptions, there might be mislabeling. For example, a host with unusual web requests that could be performed by a human might be labeled as a bot.
- Additionally, if the bots change their behaviour after deploying the IDS. This could introduce a bias known as label shift (for further details, see this article[1]). The classifier will not adapt to these changes which will experience performance decay in real scenarios.

3.5 System design and learning

With the data in hand, we can proceed to design and train a learning-based IDS. The goal is to process and use the meaningful features that are interesting for our project.

3.5.1 Data snooping

Data snooping is a phenomenon that occurs when a model is trained with information that is not available in practice. This typically happens when the test set is employed for experimental purposes before the final evaluation. Data snooping can result in overly optimistic results, where the model appears to perform exceptionally well on its training data but fails to generalize effectively to new, unseen data.

For the sake of completeness, in the article titled "Dos and Don'ts of Machine Learning" [1], 3 types of data snooping are distinguished : **test snooping**, **temporal snooping** and **selective snooping**. The overview of these snooping types can be found in Table 8 of the article :

- **test snooping** :
  - Preparatory work : This aspect is not applicable to our project. We have ensured that the test set is used **ONLY** for the final model evaluation. No additional knowledge is gained from it during the learning process. Our feature selection was based only on the training datasets without any insights from the evaluation datasets.
  - K-fold cross-validation : Given our dedicated evaluation datasets, we do not use K-fold cross-validation in any of our algorithms.
  - Normalization : We have not used any normalization functions in our project, although we discuss this point as a potential improvement in this section.
  - Embeddings : Our neural network algorithm is trained only on training data and not on the training+test dataset.
- **temporal snooping** :
  - Time dependency : As stated earlier in the report : "Instead of focusing on individual timestamp values, we incorporated them as part of a broader context, such as a session or an aggregation between a request and a response. In some sense, we prioritized a session-level structure over a query-level one". Although, if bot behavior is specifically time-dependent (for example, bots that manifests at particular times of the day), our model will likely not be able to detect it since it doesn't consider individual timestamps for temporal analysis.
  - Aging datasets : In our case, this aspect is not applicable, as we use a provided dataset that is not publicly available and subject to aging.
- **selective snooping** :
  - Cherry-picking : We do not eliminate any data from our training datasets.
  - Survivorship bias : When parsing the datasets, we do not perform any data cleaning or filtering that could introduce survivorship bias.

In summary, while we have taken measures to prevent various forms of data snooping, we note that the only type of snooping that could potentially pose an issue for our project is **"time dependency"** due to our focus on session-level structures instead of individual timestamps.

3.5.2 Spurious correlations

Spurious correlations involve relying on apparent data relationships that occur "by chance". In the course of our project, we needed to select specific features that we believed were important for distinguishing between bot and human behavior. Then, we aggregated these features into sets that appeared meaningful for our objective of distinguishing between the two. We **intentionally exclude** features such as IP addresses or host to ensure that the model's selection process is not influenced by these factors.

3.5.3 Biased parameters selection

This pitfall involves selecting the right **hyperparameters** during training based on the testing dataset to achieve high accuracy. However, it's important to recognize that the IDS may behave differently in real-world situations.

In our project, as mentioned earlier, we initially did an aggregation of the most relevant features, guided by our assumptions and background knowledge. Then, we explored several combinations of these aggregated features to identify the model with the best performance. The model's performance is then assessed using the **test set**. However, we did not perform fine-tuning on hyperparameters. Most of our hyperparameters are left to the default ones because, in our case, knowing that we have several algorithms and that each of them has its own hyperparameters, we would have to perform a grid search on **all** the hyperparameters of **all** the algorithms, which would be very time-consuming and ressource demanding. We therefore decided to leave the hyperparameters to their default values.

In the case, we would had adjusted these settings, we need to be careful to perform the grid search on the training set and not on the test set. Indeed, if we perform the grid search on the test set, we would have a bias on the test set and the performance of the model would be overestimated.

3.6 Performance evaluation

The way that we evaluate our model can greatly influence the outcome and lead to misleading and biased results.

3.6.1 Inappropriate baseline

To show to what extent a novel method improves the state of the art, it is vital to compare it with previously proposed methods.

In our project, we decided to create 5 different models and see how they perform by comparing them side by side since there exists no universal algorithm that outperforms all the other. We will evaluate the different models not only on accuracy but also on precision, recall, F1-score and other relevant metrics such as detection rate, false alarm rate, etc.

Indeed, if we directly choose a complex learning method it will increase the chance of overfitting and will raise also other problems such as security, performance, etc. Thus, for a baseline model, we choose a straightforward algorithm namely `logistic regression`.

3.6.2 Inappropriate performance measures

When assessing a model, choosing the **right performance metric** is very important, especially for security-related applications. In our situation, where we aim to develop an IDS, relying on a single performance metric is inadequate, **particularly with imbalanced classes**. We have used accuracy (more details will be given in the next sections) as a measure, but this will not be ideal in our case as it does not accurately reflect the rate of true positive and false positive predictions. As a matter of fact, in our datasets, there are 5,474

entries for bots and 343,835 for webclients which clearly indicates imbalanced classes. As mentioned previously, we preferred metrics like **detection rate**, **false alarm rate**, **false negative**, **true negative**, **precision**, **recall** and **F1-score** that are more appropriate for our project.

3.6.3 Base-rate fallacy

The base-rate fallacy occurs when the prevalence of a class (in our case, the base rate of the negative class which are legitimate hosts) in the dataset is not correctly considered while evaluating the performance of the IDS. More precisely, if the class is predominant, even a very low false-positive rate can result in a surprisingly high numbers of false positive. This can lead to a biased perceptions of the system's effectiveness.

For example, in our dataset, the large majority of entries are legitimate webclients, as indicated by the 343,835 entries for `webclients_tcpdump.txt` compared to just 5,474 for `bots_tcpdump.txt`. In such a scenario, even if the IDS has a low false-positive rate, the scale of the volume of legitimate traffic can result in a significant number of false positives. This can be problematic as it could lead to unnecessary alarms.

Moreover, the high accuracy score that might be obtained in such an imbalanced dataset can be misleading. A model might achieve high accuracy by simply predicting the majority class (in our case webclients), but this doesn't mean it is effective in identifying actual threats (in our case bots).

The ideal IDS should balance the following:

- **Detection rate (true positive rate)** $(P(D|B))$: high probability of correctly identifying actual threats.
- **False alarm rate (false positive rate)** $(P(D|\neg B))$: low probability of mistakenly identifying legitimate traffic as a threat.
- **False negative rate** $(P(\neg D|B) = 1 - P(D|B))$: low probability of failing to detect actual threats.
- **True negative rate** $(P(\neg D|\neg B) = 1 - P(D|\neg B))$: high probability of correctly identifying legitimate traffic.

where:

- B and $\neg B$: bot or not bot behaviour
- D and $\neg D$: detection as bot or human by the IDS

In essence, an effective IDS must minimize both false positives and false negatives. This requires a good understanding of the dataset's characteristics and the selection of relevant metrics.

Considering the base-rate fallacy, it is important to use metrics like precision, recall and the F1-score in addition to accuracy. These metrics provide a more global view of the model's performance, especially in the context of imbalanced datasets.

3.7 Deployment and operation

This stage marks the final step where we deploy our solution to address the security problem in practical scenarios.

3.7.1 Lab-only evaluation

Our approach is susceptible to this pitfall. The analysis is conducted and tested in a highly controlled environment. In the real world, the landscape is far more diverse and unpredictable. Our evaluation relies on the provided datasets and real-world performance may show variations. For instance, the range of "timestamps" are limited to a little more than one-hour duration in our datasets. If bots show different behavior patterns outside this timeframe, our IDS may struggle to classify them effectively.

3.7.2 Inappropriate threat model

Our learning-based algorithm is not vulnerable to certain attacks such as adversarial preprocessing or poisoning because we used a closed dataset provided by the professor.

Nevertheless, we did make certain assumptions, such as trusting the professor's datasets. If the initial datasets were tampered with or poisoned, there is little we can do and our model could be compromised. Additionally, assumptions that we made about distinguishing humans from bots, if incorrect, may affect our model's performance.

3.8 Summary of the pitfalls

This section provides a summary of the 10 pitfalls and whether they are applicable to our situation :

1. **Sampling bias** : **Yes**, we have this pitfall.
2. **Label inaccuracy** : **No**, we do not have this pitfall.
3. **Data snooping** : **Yes**, we have this pitfall.
4. **Spurious correlations** : **No**, we do not have this pitfall.
5. **Biased parameters selection** : **No**, we do not have this pitfall.
6. **Inappropriate baseline** : **No**, we do not have this pitfall.
7. **Inappropriate performance measures** : **No**, we do not have this pitfall.
8. **Base-rate fallacy** : **Yes**, we have this pitfall.
9. **Lab-only evaluation** : **Yes**, we have this pitfall.
10. **Inappropriate threat model** : **No**, we do not have this pitfall.

4 Algorithms and techniques used

In selecting the algorithms for this project, our initial step involved determining the most suitable intrusion detection strategies to implement. The seminal work on the base-rate fallacy by Axelsson [2] highlighted 3 major types of intrusion detection : **anomaly**, **signature** and **classical**.

4.1 Our choice

We opted to focus on **anomaly detection** that is a technique that identifies deviations from expected behavior within a system or dataset. In the context of our project, it is particularly valuable because it allows us to differentiate between human-generated and bot-generated traffic without relying on predefined "signature" files. This is especially pertinent as we are not provided with such signature files in our dataset.

Anomaly detection, as a machine learning approach, is well-suited to handling cases where the behavior of bots may vary and evolve over time, making it challenging to create precise signatures for them. Instead, anomaly detection algorithms learn the baseline behavior of the network and can identify deviations from this norm. These deviations may indicate the presence of bots, which often exhibit patterns that deviate significantly from human traffic.

4.2 Baseline algorithms

We learnt several ML algorithms during the first year of our master. We chose to concentrate on the followings :

Algorithm 1 : Logistic regression

Logistic regression is often the starting point for machine learning tasks, particularly in classification problems. We used it to establish our performance baseline, using its simplicity and interpretability.

Algorithm 2 : Decision tree

Our choice of the decision tree algorithm was motivated by its simplicity, offering a simple yet effective method for classification tasks.

4.3 Other and more complex algorithms

Algorithm 3 : Random forest

After using decision trees, we moved on to random forests to potentially improve performance beyond what a single decision tree can achieve. Based on our knowledge gained in machine learning courses from last year, random forests mitigate the over-fitting problems typically associated with decision trees.

Algorithm 4 : Neural networks

Neural networks represent a substantial increase in complexity over our previous models. We wanted to explore the performance of our model with a more advanced algorithm. Neural networks are known for their ability to understand complex patterns in data, functioning as black boxes.

Algorithm 5 : KNN

We chose the K-nearest neighbor method for its simplicity and straightforward implementation.

5 Discussion

In this section, we will discuss the result we obtained using the different combination of features with different algorithms. To do so, we will use different metrics.

To be secure, an IDS must have :

- A high Bayesian detection rate ($P(B|D)$) which means that if there is a detection, it is likely a real bot.
- A high assurance that if no alarm sounds, there's no intrusion : ($P(\neg B|\neg D)$).

5.1 Definition of metrics used

1. **Precision** : This metric answers the question : "Of all the instances the model labeled as a particular class, how many were actually that class" ? It's calculated as :

$$\text{Precision} = \frac{\text{True positives}}{\text{True positives} + \text{False positives}}$$

2. **Recall** : This metric answers the question : "Of all the actual instances of a particular class, how many did the model correctly identify"? It's calculated as :

$$\text{Recall} = \frac{\text{True positives}}{\text{True positives} + \text{False negatives}}$$

3. **F1-Score** : This metric provides a balance between precision and recall. It's particularly useful when the class distribution is uneven. It is the harmonic mean of precision and recall:

$$\text{F1-Score} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

4. **Support** : This shows the number of actual instances for each class in the dataset that was used to compute these metrics.

5. **Accuracy** : This gives the overall proportion of predictions that were correct, across both classes. It's calculated as :

$$\text{Accuracy} = \frac{\text{True positives} + \text{True negatives}}{\text{True positives} + \text{False positives} + \text{False negatives} + \text{True negatives}}$$

6. **Macro avg** : This averages the unweighted mean per label. In the code, precision, recall and F1-score are each averaged, giving equal weight to both classes. It is particularly informative as there's a class imbalance, as it treats both classes equally.
7. **Weighted avg** : This averages the support-weighted mean per label. Metrics are calculated by considering the number of true instances for each label (support). This gives a weighted average which is more representative when there's a class imbalance.
8. **Bayesian Detection Rate** : This metric is used in probabilistic models to measure the security of a classifier in identifying true instances of a class, given the probability of that class. It uses bayesian probability to adjust the detection rate based on assumed distributions of the classes. It's calculated as :

$$\text{BDR} = \frac{(\text{True positive rate}) \times (\text{Probability of bot})}{(\text{True positive rate}) \times (\text{Probability of bot}) + (\text{False positive rate}) \times (1 - \text{Probability of bot})}$$

We found on a scientific article, an excellent explanation that argues why accuracy is a biased metric if we want to build a **secure** IDS : "Accuracy is seeing frequent usage and conceptually it makes sense. However, formulaically, its reliance on TN in the numerator, giving it equal weight to TP, is invalid in cybersecurity due to the high volume of TNs relative to TPs and their nonexistent value. It's FPs that are a concern, in volume, while TNs are completely irrelevant. Essentially, the issue is a deviation between a theoretic assessment of the technique and a real-world assessment. From a purely theoretical point of view, accuracy appears to be a relevant and completely applicable metric. From the **real world**, cybersecurity point of view, **accuracy** is completely **irrelevant**. The problem being that accuracy assumes all samples are of **equal weight**. Since that isn't the case, **accuracy cannot be used**." [3]

For instance, consider a scenario with an accuracy of 84.16%. At first glance, this may appear as an effective model. However, this is misleading. For instance, the number of true positives is **only 2 out of 12**. In the context of cybersecurity and IDS, this is problematic. A secure IDS must identify as many bots as possible to be considered efficient and the model in question fails to meet this criterion. Its accuracy is disproportionately influenced by the high number of true negatives, a result of the larger sample size for human users. Thus, while the accuracy appears high, it's actually biased and not indicative of the model's effectiveness in detecting bots.

Listing 4: Scenario demonstrating that accuracy is irrelevant

```
####
False alarm rate and detection rate...
Detection rate : 16.666666666666664 %
False alarm rate : 8.333333333333332 %
False negative rate : 83.33333333333334 %
True negative rate : 91.66666666666666 %
Accuracy : 84.16666666666667 %
####
Total host :  120
```

```
Number of true positive :  2
Number of false positive :  9
Number of false negative :  10
Number of true negative :  99
####
```

$$\text{Accuracy} = \frac{2+99}{2+9+10+99} = 0.8416666666666667$$

5.2 Eval2 : evaluation dataset

As a reminder, in this dataset, hosts can emit traffic from a human and from a bot.

Our results on this dataset indicate poor performance. We believe that it is totally normal and that the core of the issue lies in the lack of training data for the "human+bot" category. Machine learning models learn to make predictions based on the data they are trained on. Without representative examples of "human+bot" behavior, the model is essentially "blind" to this category.

There are also several other issues to achieve a good performance for this dataset :

- **Nature of "human+bot" behavior** : it represents a hybrid of legitimate user behavior and malicious bot activities. This duality makes it more complex and less predictable than the distinct "human" or "bot" categories. These behaviors can intermingle in ways that aren't obvious, creating a mix that's difficult to categorize without specific examples.
- **Imbalanced training data** : As explained earlier, with 345k entries for "human" and only 5k for "bot", the model is heavily biased in favor of recognizing human behaviors. This imbalance exacerbates the challenge, as the model is more likely to misclassify "human+bot" instances as purely "human" due to its learning bias.

It's important to note that even with all these issues, that does not mean that we did not try to provide a 'solution' (even if it is not satisfactory). For more information about the solution, see the threshold section.

5.3 Diagrams

For the next diagrams, we selected the `logistic regression` algorithm as a representative choice for generating diagrams due to its simplicity and effectiveness for binary classification tasks. It serves as an illustrative example of how our approach works. Additionally, in the section 5.3.3, we will also give a comparison of the algorithms we used with a fixed set of features, namely "Combination 1", for the following metrics : precision, recall and F1-score.

5.3.1 Metrics depending on combination of features

In this subsection, we will conduct a comparative analysis on different metrics in relation to our feature sets to identify which feature sets gives the best results. We will also describe the results we obtain. Our approach will begin with presenting the results from our test with `eval1` and then we will move on to explore and discuss the findings from our test with `eval2`.

For eval1

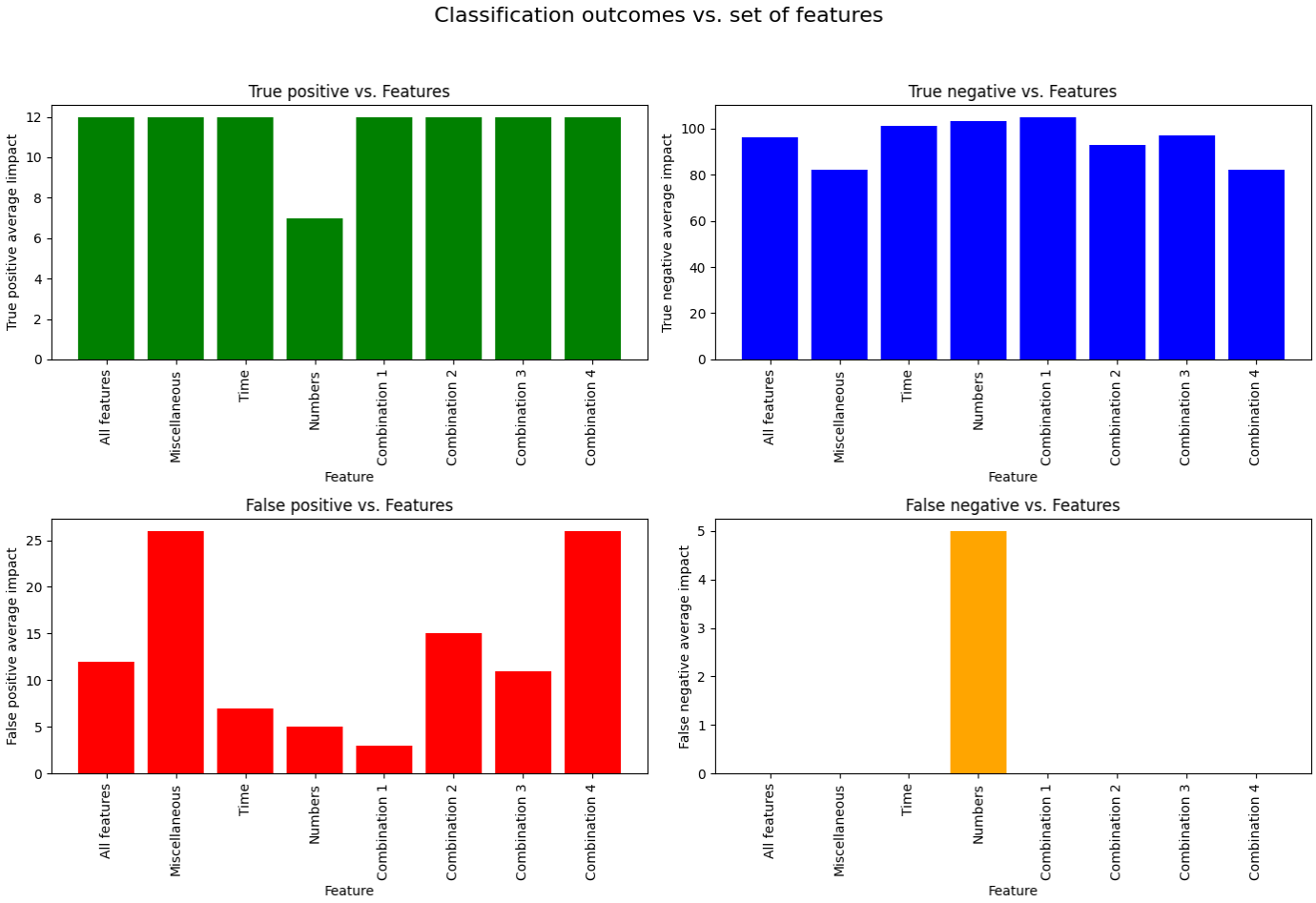

Figure 2: Classification outcomes vs. set of features on eval 1

The diagram shows the evaluation of logistic regression against 4 different metrics, namely true positive, false positive, true negative and false negative, on the evaluation set `eval1`. Each set of features was used to train the model and, only after that, it was tested on the `eval1` dataset.

We can observe the following results :

- **True positives (TP)** : The model correctly identified bots 12 times for all feature sets, except when using the "Numbers" feature, where it identified 7.
- **False positives (FP)** : The model incorrectly identified webclients as bots 12 times when all features were used, peaking at 26 times with "Miscellaneous" and "Combination 4". The lowest FP count was 3 with "Combination 1".
- **False negatives (FN)** : The model did not have any false negatives, except when using the "Numbers" feature alone, which resulted in 5.
- **True negatives (TN)** : The model correctly identified webclients as humans most often when using "Combination 1" (105 times), and least often with "Miscellaneous" and "Combination 4" (82 times).

The **green bars (TP)** are consistently high except for "Numbers", indicating that all feature sets except that one are good at identifying bots. The **blue bars (TN)** shows slight variations, with "Combination 1" performing best in correctly identifying humans. The **red bars (FP)** show more variation, indicating that "Miscellaneous" and "Combination 4" lead to more incorrect bot classifications, while "Combination 1" is better at avoiding these errors. Lastly, the **yellow bar (FN)** is nearly nonexistent, showing that the model rarely misses a bot classification, except for the "Numbers" feature, which underperformed compared to the others.

In summary, the model is quite effective, especially when using **"Combination 1"** which has the highest TN and lowest FP rates, suggesting a good balance of precision and recall for this particular feature set.

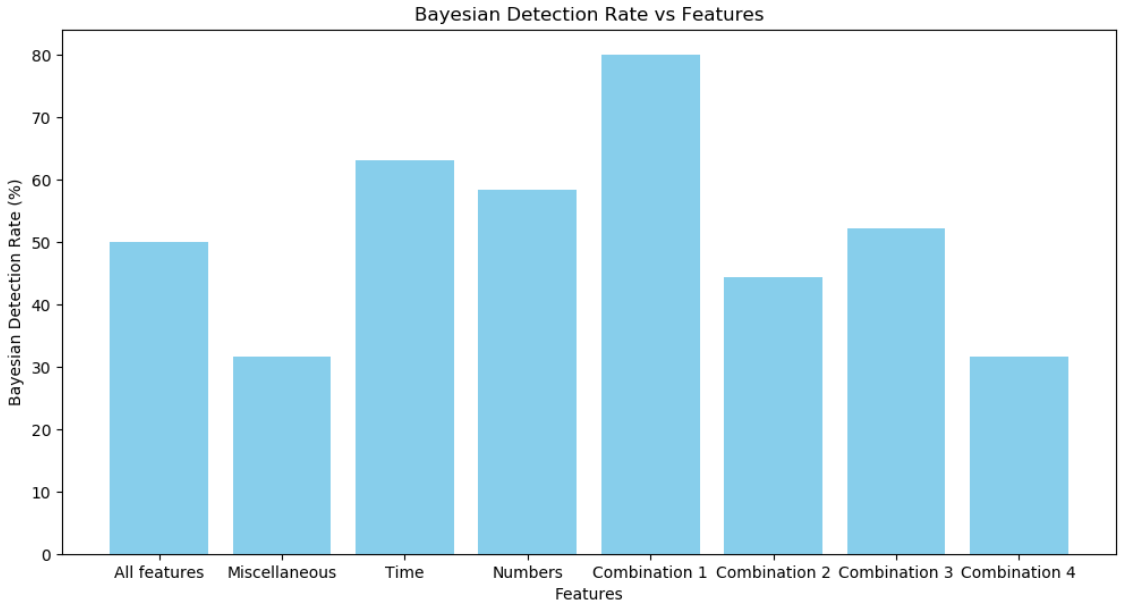

Figure 3: Bayesian detection rate vs. set of features on eval 1

The Figure 3 plots the Bayesian detection rate against feature sets using a logistic regression model on the **eval1** evaluation set.

We can observe the following results :

- **Feature set performance** : The feature set labeled "Combination 1" has the highest Bayesian detection rate, which suggests that the particular combination of features in this set is most effective for classifying hosts as bots or humans within the evaluation set `eval1`.
- **All features vs. selective features** : Interestingly, using all features does not result in the highest detection rate. This may indicate that some features are not contributing positively to the model's performance and could, therefore, be introducing noise or overfitting issues.
- **Time and numbers** : The "Time" and "Numbers" feature sets individually offer pretty good detection rates, suggesting that these categories of features hold some predictive value on their own.
- **Miscellaneous features** : This set appears to be the least effective on its own. This could mean that the features grouped under "Miscellaneous" may not be strong indicators for detecting bots or may require to be used in conjunction with other features to provide better results.
- **Combination feature sets** : "Combination 2", "Combination 3" and "Combination 4" have lower detection rates than "Combination 1", but they still outperform the "Miscellaneous" sets. "Combination 4" has the lowest performance among the combination sets, which might indicate that it includes features that degrade the model's performance.

In summary, "Combination 1" seems to be the most effective set of features for the model in detecting bots in the evaluation set `eval1`. A higher Bayesian detection rate implies a more reliable IDS, leading to fewer false positives and negatives, which is critical in security.

For eval2

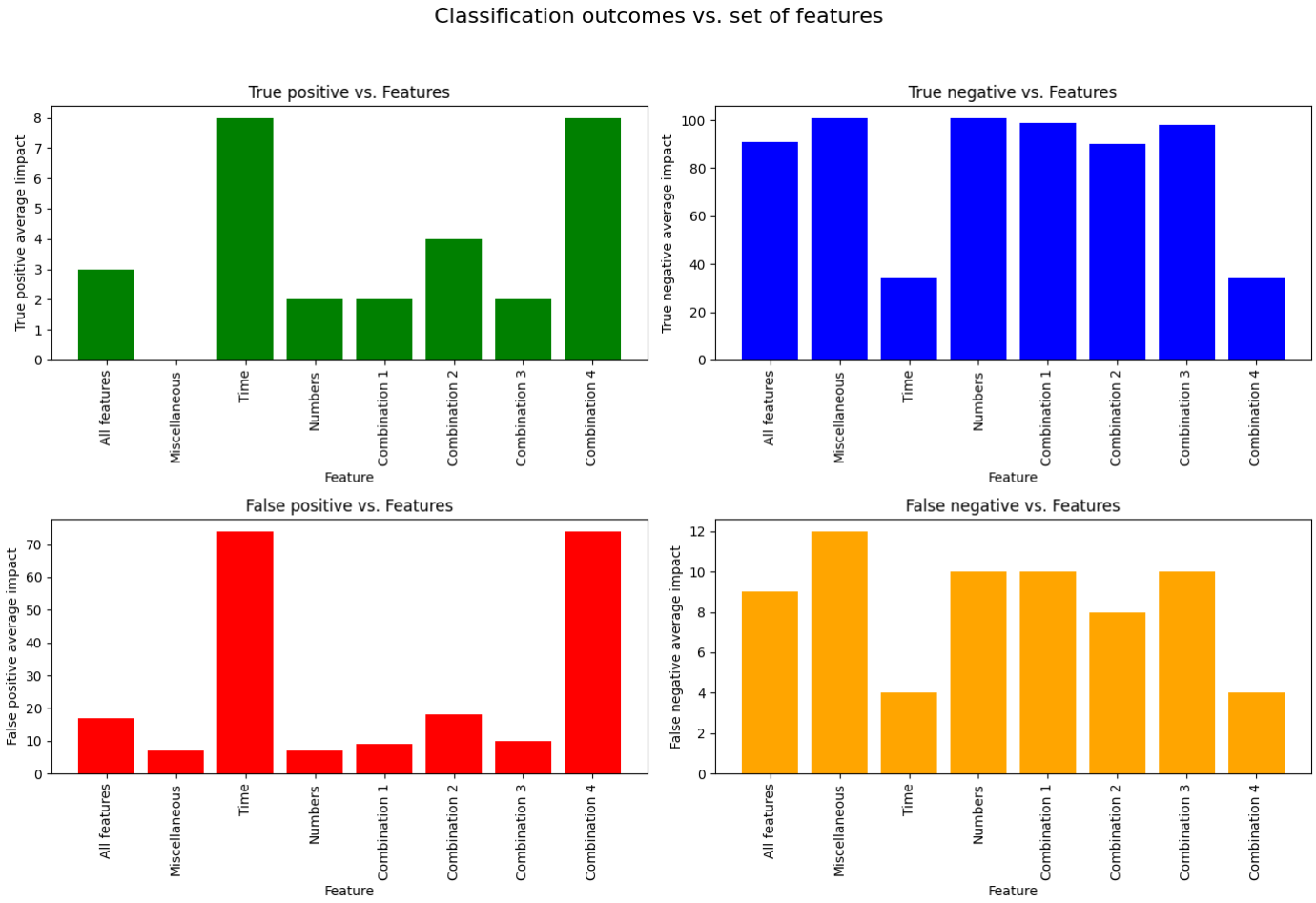

Figure 4: Classification outcomes vs. set of features on eval 2

On this evaluation dataset, the performance is extremely different (in a bad way).

We can observe the following results :

- **True positives (TP)** : Contrary to `eval1`, the number of true positive is extremely low. The best sets of features (”Time” and ”Combination 4”) identified only 8 bots while the worst set of feature for `eval1` identified 7 bots.
- **False positives (FP)** : Contrary to `eval1`, the false positives are more present in general, with the worse outcome when considering ”Time” and ”Combination 4” where they have more than 70 false positives each. The lowest are ”Miscellaneous” and ”Numbers” with less than 10.
- **False negatives (FN)** : The false negative are present for every set of features compared to `eval1` where only the ”Numbers” feature was impacted. The ”Miscellaneous” feature is even worse since it did not find any bot from the evaluation set. The ”Time” and ”Combination 4” have the lowest number of false negative.

The model did not have any false negatives, except when using the "Numbers" feature alone, which resulted in 5.

- **True negatives (TN)** : Contrary to `eval1`, the true negatives are also impacted but not as much as other metrics. Again, "Time" and "Combination 4" behave the most differently from `eval1`.

The **green bars (TP)** are consistently low except for "Time" and "Combination 4". The **blue bars (TN)** shows slight variations with `eval1` but, for "Time" and "Combination 4", it is disastrous, there is very few true negatives. The **red bars (FP)** show a bit of variation, but again for "Time" and "Combination 4", the results are really the worst. Lastly, the **yellow bar (FN)**, while almost nearly nonexistent in `eval1`, in this case, every sets of features has at least 4 false negatives showing that the model misses a lot of a bot classification.

In summary, the model is really not effective and no set of feature is actually outstanding. It is not surprising since the model was not trained with a dedicated dataset containing hosts acting as webclients and bots.

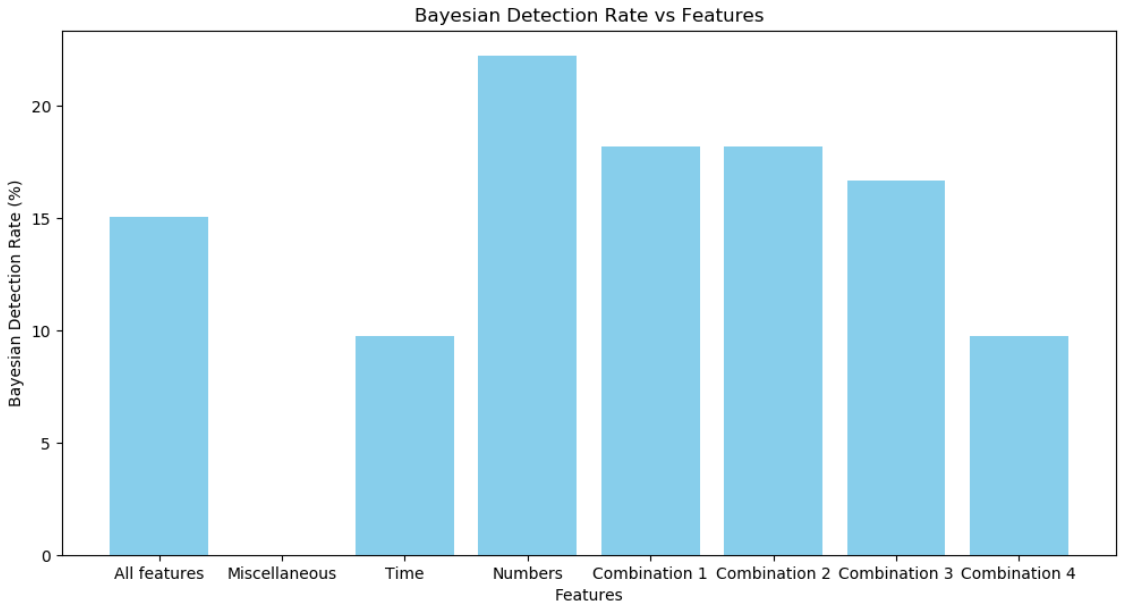

Figure 5: Bayesian detection rate vs. set of features on eval 2

The Figure 5 plots the Bayesian detection rate against feature sets using a logistic regression model on the **eval2** evaluation set.

Again, we can directly see that we have really poor results compared to the `eval1` evaluation dataset.

We can observe the following results :

- **Feature set performance** : The feature set labeled "Numbers" gives the highest Bayesian detection rate, which is pretty counter-intuitive as the best sets of features for `eval1` was "Combination 1".
- **All sets of features** : As a general rule, compared to the eval dataset, the Bayesian detection rate is extremely low. The best rate in `eval2` is more or less 22 while the worst rate in `eval1` was more or less 30.
- **Numbers** : "Numbers" feature set individually offer the best Bayesian rate.
- **Miscellaneous features** : This set is the worst. Even on `eval1`, it was the least effective. This could mean that the features grouped under "Miscellaneous" may not be strong indicators for detecting bots or may require to be used in conjunction with other features to provide value.
- **Combination feature sets** : Contrary to **eval1**, none of the combinations outperform the "Numbers" feature set. However, "Combination 4" is the least effective among the combination feature set as seen in **eval1**. It could mean that some features in this combination degrade the model's performance.

In summary, none of the sets of features achieve a good performance with the highest rate reaching only more than 20% for the "Numbers" feature. This outcome is coherent since it shows a relatively low false positive rate, associated with a number of true positives of just 2 which indicates a low detection rate. However, its highest result in terms of BDR is mainly due to the fact that it has fewer false positives than the other feature sets which gives it a slightly higher performance.

5.3.2 Accuracy VS. false alarm rate

As precised in the introduction of this section, we used the logistic regression for the next diagrams.

For eval1

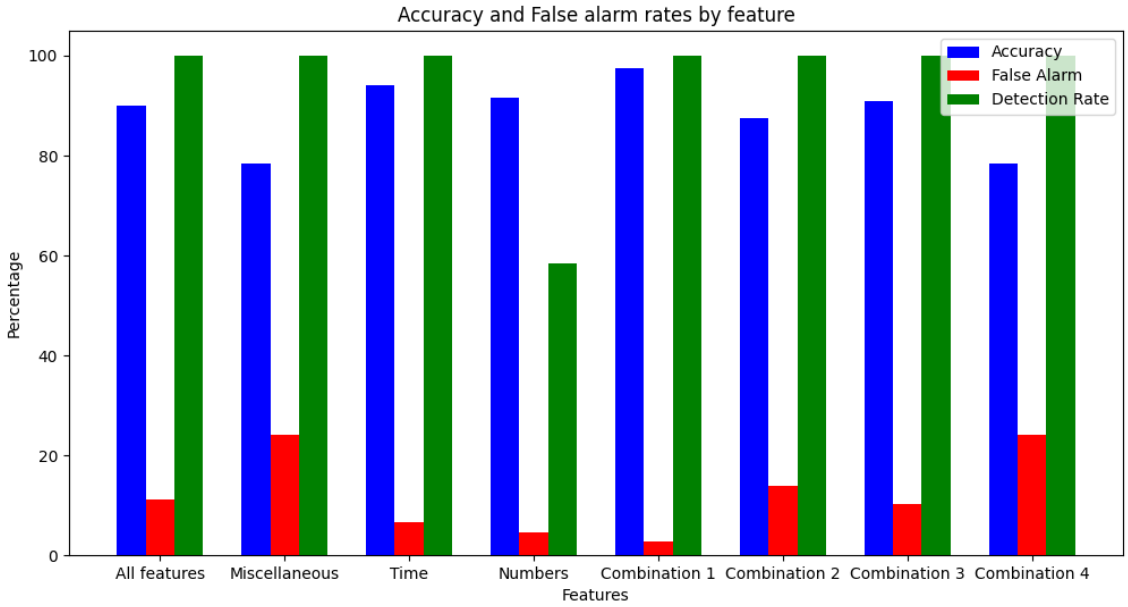

Figure 6: Accuracy and false alarm rate by features on eval1

- **All features (Accuracy: 90%, False Alarm: 11.11%, Detection: 100%)** : This set provides high accuracy and detection rate but has a moderate false alarm rate. It suggests a good balance between detecting bots and avoiding misclassifying humans.
- **Miscellaneous (Accuracy: 78.3%, False Alarm: 24.07%, Detection: 100%)** : Lower accuracy and higher false alarm rate indicate that relying only on this set leads to more false positives while still detecting all bots.
- **Time (Accuracy: 94.16%, False Alarm: 6.48%, Detection: 100%)** : High accuracy with a low false alarm rate suggests that time-based features are highly effective in distinguishing bots from humans without many false positives.
- **Numbers (Accuracy: 91.66%, False Alarm: 4.63%, Detection: 58.33%)** : While this set has high accuracy and a low false alarm rate, the detection rate is significantly lower, indicating that it misses many bots.
- **Combination 1 (Accuracy: 97.5%, False Alarm: 2.77%, Detection: 100%)** : The best performing set with high accuracy, low false alarm and perfect detection rate, indicating an very good balance of features.
- **Combination 2 (Accuracy: 87.5%, False Alarm: 13.88%, Detection: 100%)** : This set of features has a moderate performance with a relatively high false alarm rate.
- **Combination 3 (Accuracy: 90.83%, False Alarm: 10.18%, Detection: 100%)** : Similar to the "All features" set, but with a slightly lower false alarm rate.
- **Combination 4 (Accuracy: 78.33%, False Alarm: 24.07%, Detection: 100%)**

: Similar to "Miscellaneous", it has lower accuracy and higher false alarms.

In summary, "Combination 1" is the most effective set of features, having a high accuracy, few false alarms and very good detection capabilities. On the other hand, "All features" and "Combination 3" are also effective, however, they have slightly higher false alarm rates. "Numbers" feature set, despite its high accuracy, is less effective in detecting bots. For the remaining features, namely "Miscellaneous" and "Combination 4", they have lower accuracy and higher false alarms, making them less suitable.

For eval2

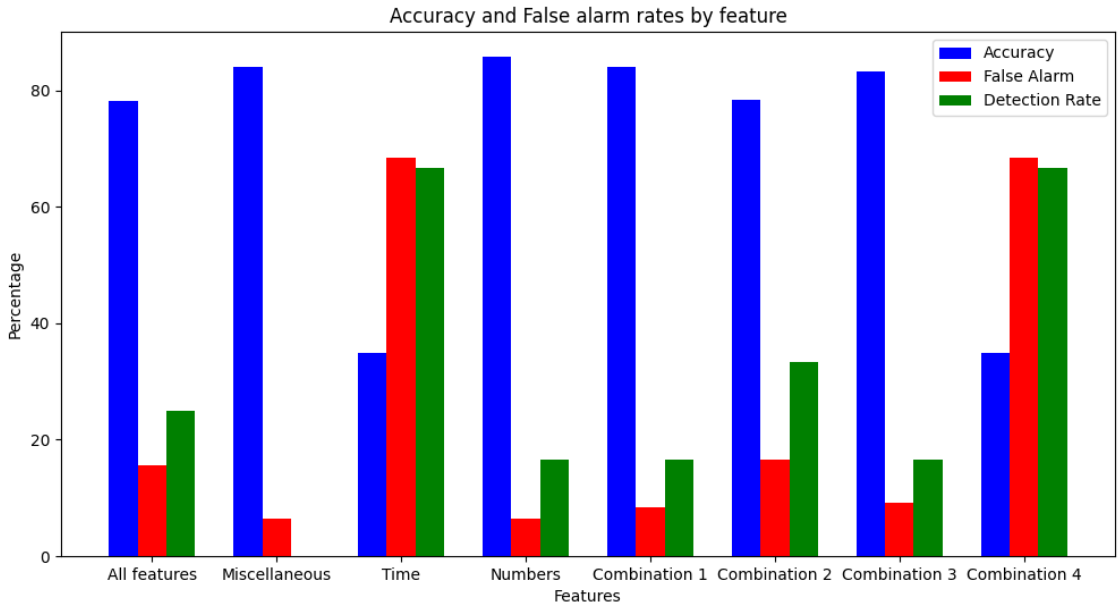

Figure 7: Accuracy and false alarm rate by features on eval 2

Observing the results, it's evident that while the accuracy remains relatively stable, the detection rate has significantly dropped across most categories, with the exceptions being "time" and "Combination 4".

- **All features (Accuracy: 78.3%, False Alarm: 15.7%, Detection: 25%)** : This set shows a significant decrease in accuracy and detection rate compared to `eval1`, with a slight increase in false alarm rate. It indicates a significant decrease in the model's performance.
- **Miscellaneous (Accuracy: 84.16%, False Alarm: 6.48%, Detection: 0%)** : Despite a higher accuracy and a lower false alarm rate than `eval1`, this set fails to detect any bots, indicating a major issue in its ability to identify threats in `eval2`.
- **Time (Accuracy: 35.0%, False Alarm: 68.52%, Detection: 66.67%)** : This feature set shows a drastic reduction in accuracy and a significant increase in false alarm

rate, although the detection rate is moderate. It suggests that time-based features are not reliable in this evaluation scenario.

- **Numbers (Accuracy: 85.83%, False Alarm: 6.48%, Detection: 16.67%)** : Improved accuracy and low false alarm rate are seen here, but the detection rate is quite low, indicating that many bots are not identified.
- **Combination 1 (Accuracy: 84.17%, False Alarm: 8.33%, Detection: 16.67%)** : While this set maintains high accuracy and a relatively low false alarm rate, the detection rate is significantly lower compared to `eval1`, showing diminished effectiveness in identifying bots.
- **Combination 2 (Accuracy: 78.33%, False Alarm: 16.67%, Detection: 33.33%)** : This set shows a modest accuracy with an increased false alarm rate and moderate detection rate, suggesting an overall drop in performance.
- **Combination 3 (Accuracy: 83.33%, False Alarm: 9.26%, Detection: 16.67%)** : Despite maintaining a good accuracy and a moderate false alarm rate, the low detection rate indicates a significant issue in identifying bots in `eval2`.
- **Combination 4 (Accuracy: 35.0%, False Alarm: 68.52%, Detection: 66.67%)** : Similar to the "Time" feature set, this combination shows poor accuracy and a high false alarm rate, though with a moderate detection rate. This indicates a lack of reliability in this scenario.

In summary, all the features shows a notable drop for the detection rate. The "Miscellaneous" set, despite higher accuracy and lower false alarms, completely fails to detect any bots. "Combination 1", while still showing high accuracy, experiences a significant drop in detection capabilities. "All features", "Combination 2", and "Combination 3" also show decreased effectiveness, especially in detection rates. The "Time" and "Combination 4" feature sets perform poorly in `eval2`, with high false alarms and low accuracy, although their detection rates are moderate.

Conclusion : An objective of the analysis of these diagrams was to challenge our initial assumptions (before this project and reading the articles), particularly regarding the reliability of accuracy as a metric for evaluating a model's security. Our results demonstrate that **accuracy alone is not a sufficient and reliable indicator** when it comes to assessing the robustness of a model in terms of security. It underlines the need for a more nuanced approach to assessing model performance, particularly in the context of security applications.

5.3.3 Comparison between the algorithms based on metrics

Concerning the following diagrams, we had to choose a set of features, as plotting every set was not possible. Our choice, "Combination 1", was driven by logic and performance considerations, making it the most relevant for comparing various algorithms.

Consistent with our approach, we created distinct diagrams for `eval1` and `eval2`. Additionally, in this instance, we differentiated between bots and humans across three different

metrics : **precision**, **recall** and **F1-score**.

For eval1

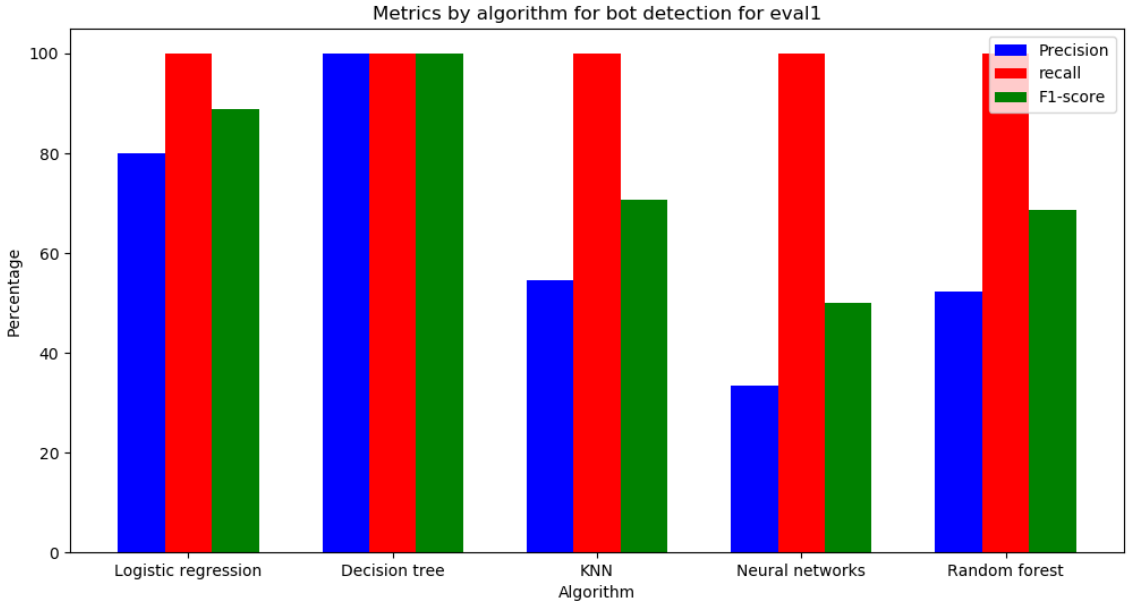

Figure 8: For bot

Each algorithm achieved a 100% recall rate. This implies that all bots in the `eval1` set were correctly identified as bots by each model.

The precision rates vary significantly among the models. The decision tree algorithm stands out with a 100% precision score, suggesting that every host it classified as a bot was indeed a bot. On the other hand, neural networks had the lowest precision at 33.33%, indicating that many hosts classified as bots were not actually bots (false positive). KNN has also a lower precision, accompanied by lower F1-scores. It indicates less reliability in classifying hosts as bots compared to other models.

In conclusion, the decision tree algorithm appears to be the most effective for classifying bots in the `eval1` evaluation set, achieving perfect scores across for all metrics.

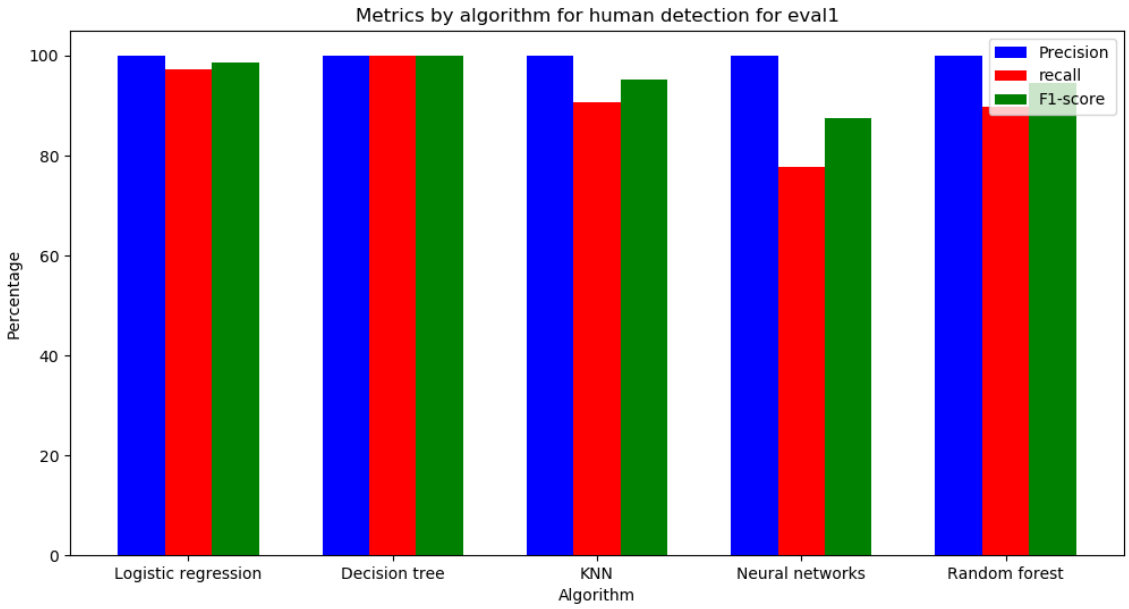

Figure 9: For human

Each algorithm achieved a 100% precision rate, indicating that every host classified as human was indeed a human. This is an excellent result, showing a high degree of accuracy in classifying humans. However, there is variability in recall rates, with the decision tree model achieving perfect recall, while neural networks have the lowest recall at 77.77%. As a reminder, recall measures how well the model identifies all actual humans in the dataset. The F1-score follow a similar pattern, with the decision tree model being perfect and neural networks having the lowest score.

The decision tree algorithm achieved perfect scores in all metrics, making it highly effective in identifying human hosts in the `eval1` set. The neural networks model, despite perfect precision, has significantly lower recall and F1-score, indicating a tendency to miss classifying some humans correctly.

In conclusion, the decision tree model show exceptional performance in identifying human hosts, achieving perfect scores across all metrics. The neural networks model, although accurate in its positive predictions (precision), falls short in correctly identifying all human hosts, as reflected in its lower recall and F1-score.

For eval2

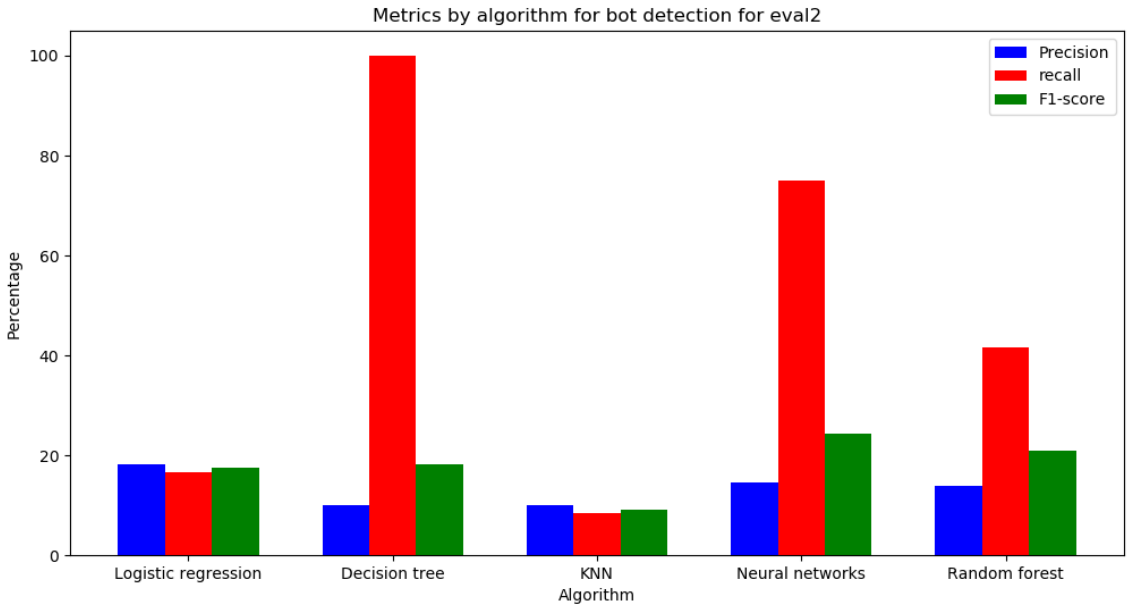

Figure 10: For bot

Compared to the `eval1` set, all models show very poor performance in the `eval2` set. This is due to the human+bot behavior, since the model was only trained using data containing either bot behavior or human behavior, not both of them.

The precision scores are considerably low for all models, indicating a high rate of false positives (hosts incorrectly classified as bots). The recall rates vary significantly. The decision tree model has a 100% recall, identifying all actual bots, but at the cost of a very low precision. Other models like KNN show both low precision and recall, indicating overall poor performance. The F1-scores are also generally low, with the neural network model showing the highest score (24.32%). However, this is still quite low, suggesting a lack of balance between precision and recall.

The decision tree model, while achieving a 100% recall, has a very low precision (10%), leading to many false positives. This indicates an inability to generalize well to the `eval2` dataset.

The overall low performance across all models on the `eval2` is not surprising since, as explained before, it contains mixed host behavior showing at a certain time bot behavior and at an other time human behavior. This introduces features associated with both human and bot behavior, which are new to the models.

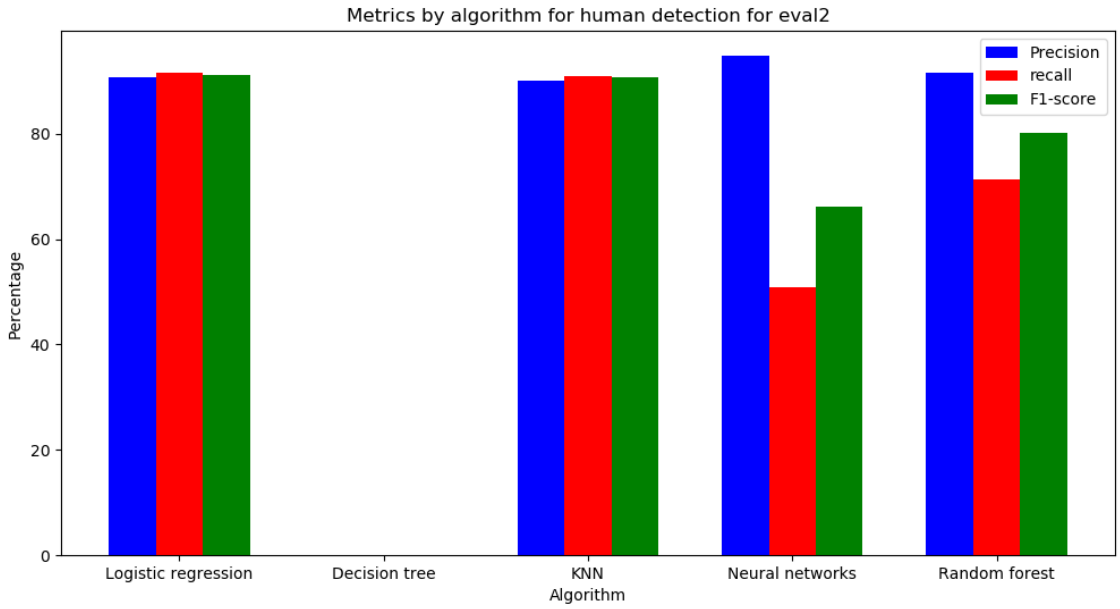

Figure 11: For human

The decision tree algorithm, which performed excellently in the `eval1` set, has a complete drop in performance (0% across all metrics) in the `eval2` set which means that all hosts are classified as bots. Others models show pretty good precision and recall (compared to bots metrics), indicating effective classification of human hosts, though with some degree of error compared to the `eval1` results.

While the precision is high, the recall is notably lower, indicating that while most hosts classified as humans are correctly identified, a substantial proportion of human hosts are being missed. The F1-scores suggest that logistic regression and KNN are relatively balanced in terms of precision and recall. The neural networks model, despite its high precision, has a significantly lower F1-score due to its lower recall.

Logistic regression, KNN and random forest show good but not perfect performance, suggesting they are more robust to changes in the dataset compared to the decision tree model. Neural networks stand out for high precision but lower recall, indicating a tendency to be overly cautious in classifying a host as human.

In conclusion, the decision tree model, which were the best in the `eval1` set, failed completely in the `eval2` set.

Final conclusion :

For the `eval1` set, the decision tree algorithm demonstrated outstanding performance in identifying both bots and humans, achieving perfect scores across all metrics. While other models like neural networks showed high precision, they struggled with lower recall and F1-scores, especially in identifying humans.

On the other hand, in the `eval2` set, with mixed bot and human behavior, all models showed significantly reduced efficacy. The decision tree model, despite its excellent performance on `eval1`, showed a complete drop in performance for `eval2`. In contrast, models like logistic regression, KNN and neural networks displayed relatively better precision but struggled with recall, especially in classifying bots.

Overall, the decision tree model proved best in the less complex `eval1` set, but failed to adapt to the complexities of the `eval2` set. In contrast, logistic regression, KNN and neural networks showed some robustness to change for the human classification, indicating their potential effectiveness in more complex scenarios. The results underline the importance of selecting algorithms according to the specific characteristics and complexities of the dataset in question.

5.3.4 Visualisation of the data using t-SNE

We thought that it would be a great idea to visualize the data distribution using t-SNE, analyzing 3 feature sets : "All Features", "Combination 1" and "Numbers". The reader is reminded that t-SNE serves as a dimensionality reduction and visualization tool, not as an evaluation or algorithmic processing tool.

In our visual representations, data points relating to humans are marked in **blue**, while those representing bots are in **red**.

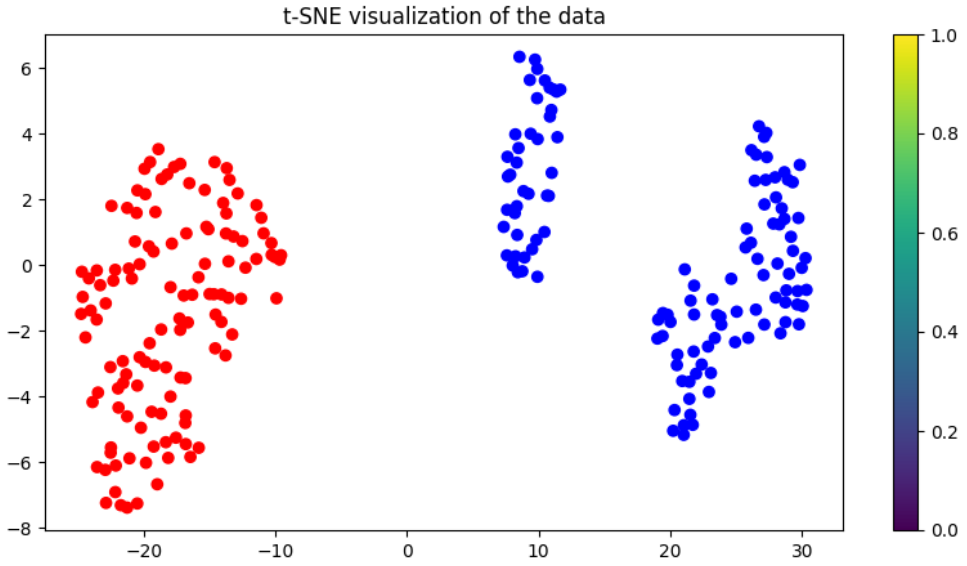

Figure 12: Visualization of training dataset for all features

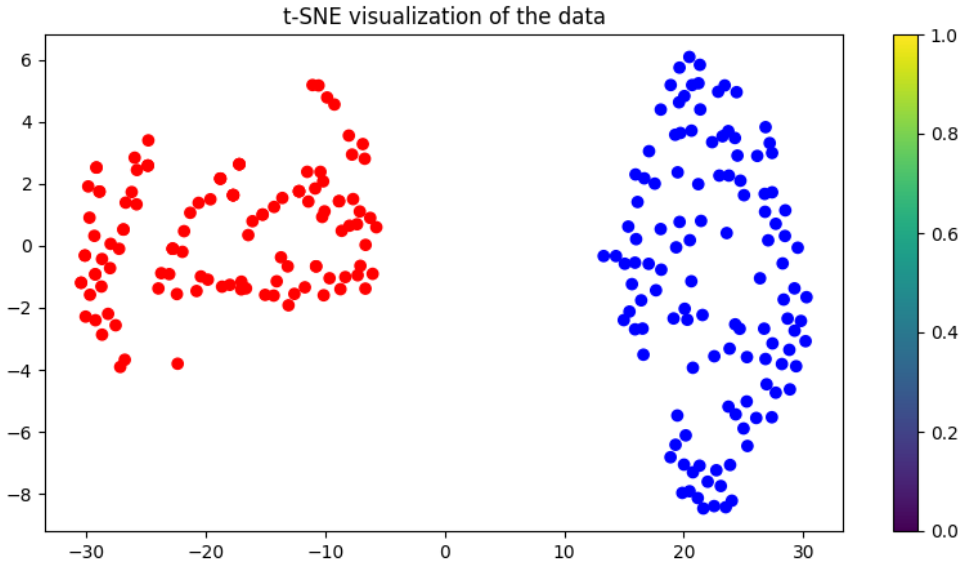

Figure 13: Visualization of training dataset for COMBI1

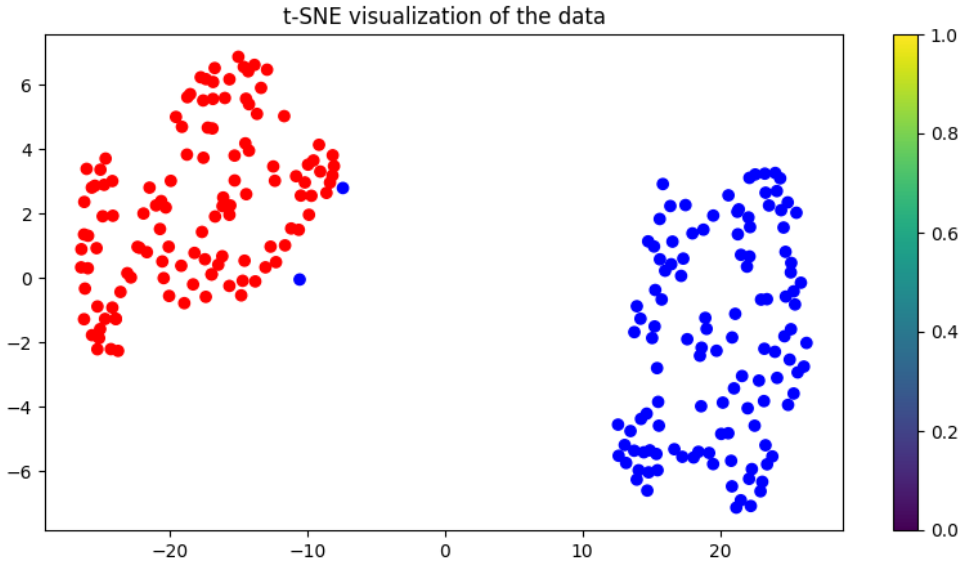

Figure 14: Visualization of training dataset for combinations of numbers

In Figure 12, we can clearly see 2 different clusters but the human data appears to be subdivided into two smaller groups. Despite this, the clusters are pretty tight. This subdivision might be attributed to the diverse spectrum of human online behaviors, which are more

complexly represented in the set of all features.

Figure 13 displays a clear demarcation between the red and blue clusters. However, the blue cluster is not as tightly formed as seen in the all-features visualization. This could indicate that the "Combination 1" features produce a good but slightly more diffuse separation of categories compared to the features used in Figure 12.

In Figure 14, we observe an interesting pattern : a mostly red cluster with a few blue points and a separate all-blue cluster. This could indicate a higher incidence of false positives, where human behavior is incorrectly classified as bot, particularly when only numerical features are considered.

5.4 Threshold for distinguishing human+bot from bot

Distinguishing between pure bot traffic and traffic generated by a combination of both bots and humans was a tough task in the project. Our classifier is designed to perform binary classification, labeling data points as either bots or humans. However, it also provides a probability score indicating its confidence in each classification. To address the complexity of mixed behaviors, we introduced a post-processing step that involves comparing the absolute difference between the two probability values assigned by the classifier.

The threshold we set for this difference is at **0.5**. When the absolute difference between the probabilities for bot and human classifications exceeds this threshold, it signifies that the model is 'certain' about its decision. It suggests that the features' results are really distinct, making it easy to definitively categorize the traffic as either purely bot or purely human.

Example for clarity

The classifier provides two probability values : the first indicating the likelihood that the host is a bot and the second indicating the likelihood that the host is human :

```
unamur 25 :  [0.71725976 0.28274024]
```

In this case, the absolute difference between the two probabilities is smaller than the threshold of 0.5, indicating that the classifier is uncertain about its classification. This outcome suggests mixed behavior within the traffic data.

In contrast, in the following example :

```
unamur 12 :  human+bot :  [0.99999 0.00001]
```

Here, the absolute difference between the probabilities is significantly greater than the threshold, indicating a higher level of confidence in the classifier's decision. In this scenario, the model is more certain that the traffic is predominantly driven by a single behavior type.

5.5 Potential improvements

We would like to reflect on our methodology and identify potential areas for improvement. While our approach was carefully considered and implemented, there are aspects that could

be improved.

First, regarding feature selection, we initially organized our features logically based on our domain knowledge. However, we later discovered that scikit-learn provides a feature selection method known as SelectKBest, which automatically identifies the most relevant features for a given task. When we applied SelectKBest to our data to determine the five best features, the results were as follows :

Listing 5: Results of SelectKBest feature selection

```
Model Accuracy: 1.0
Selected Feature Indices: [0 2 5 6 7]
Selected Feature Names: [
        'average_of_request_length_encoded',
        'type_of_requests_queried_by_hosts_encoded',
        'average_time_between_requests_encoded',
        'frequency_of_repeated_requests_in_a_short_time_frame_encoded',
        'average_number_of_dots_in_a_domain_encoded'
                        ]
```

While these features performed well on our training data, relying only on these features would lead to overfitting since, when testing them on the evaluation datasets, the results were not that great. Consequently, we opted not to use these features for our final model. Nevertheless, for the sake of completeness, the SelectKBest function is still in the code and the resulting combination of features is still available in the 'constants.py' file under the name 'LIST_OF_FEATURES_AFTER_FEATURE_SELECTION.'

Another aspect worth considering for improvement is the normalization of datasets to address the issue of unbalanced classes, as discussed in the section on the base-rate fallacy. One potential solution we found was generating 'synthetic' samples based on the bot dataset. During our research on the internet, we came across the **S**ynthetic **M**inority **O**ver-sampling **Te**chnique, a method for creating synthetic samples in machine learning to balance imbalanced datasets. Implementing SMOTE could have maybe improved the overall performance of our classifier, especially in dealing with the class imbalance issue.

Finally, something that we would have done if we had more time is to create diagrams comparing precision, recall and F1-score across **different feature sets** (although, we did compare those metrics but only between the different algorithms). The development of these diagrams could have provided clearer indications of how each feature set influences performance measures, possibly revealing underlying patterns and interactions. These visualizations could have offered a more intuitive understanding of the strengths and weaknesses of each feature set, allowing us to improve and refine our approach in the future.

5.6 Issues faced during the project

- **Handling large volumes of data** : The datasets contained a massive volume of DNS traffic, making data preprocessing and feature selection a time-consuming process.

- **Parameter tuning** : The project involved an array of parameters that required fine-tuning. This included experimenting with various combinations of features, selecting appropriate machine learning algorithms and optimizing hyperparameters for these algorithms. Choosing the right balance between different parameters while maintaining acceptable performance metrics was a pretty hard process.
- **Feature selection** : We had a lot of features to choose from, each potentially contributing to the classification task. The challenge was to identify the most relevant features that could accurately differentiate bot behavior from human behavior.
- **Avoiding pitfalls** : The awareness of potential pitfalls in machine learning, such as sampling bias, data snooping and spurious correlations, added complexity to the project (but we recognize that it was necessary to have a robust classifier). It was essential to design and implement strategies to mitigate these pitfalls and ensure the robustness and validity of the machine learning model.
- **Evaluating model performance** : Determining the real effectiveness of the IDS model required careful consideration of performance metrics. Selecting the right evaluation criteria and ensuring that the model performs well in the evaluation datasets presented its own set of challenges.

6 Conclusion

During the project, we realised that developing an IDS based on machine learning is a really complex task. We tried to design and develop an IDS that not only achieves high accuracy but also ensures security, by relying on different metrics (TP, FP, TN, FN) in distinguishing bot traffic from human traffic. It has allowed us to develop a better understanding of the pitfalls, especially the base-rate fallacy and we tried to take measures to avoid them (whenever possible).

Also, during the project, we learnt the importance of the Bayesian detection rate and his implications for the security of our learning-based IDS.

Our key takeaways and reflections about this project include :

- **Bayesian detection rate** : During the project, we learnt the importance of the Bayesian detection rate and his implications for the security of our learning-based IDS. This approach was needed in addressing the challenges posed by the base-rate fallacy. By integrating it, we could more accurately weigh the probabilities of false positives and negatives, which are critical in our IDS. This provided a better understanding of the reliability and limitations of our classifier in the different evaluation datasets.
- **Complexity of ML IDS** : Creating an effective IDS using ML involves a lot of parameters, from feature selection to algorithm choice. Achieving a good balance between these parameters while maintaining performance metrics is pretty hard.
- **Pitfall awareness** : We gained a better understanding of the pitfalls associated with machine learning presented on this article[1], particularly in the context of IDS devel-

opment.

- **Data limitations** : By recognizing the constraints of the provided datasets, such as limited data volume and specific botnet behaviors, we were able to gain a better understanding of the results generated by the model. We addressed these limitations by adapting our analysis and modeling strategies.

In conclusion, this project has been both enjoyable and very challenging. We learnt a lot of complexities involved in machine-learning, it also enhanced our understanding of the pitfalls and we now have a better perspective on developing **secure** and **accurate** classifiers.

7 Bibliography

References

[1] Daniel Arp et al. "Dos and Don'ts of Machine Learning in Computer Security". In: *CoRR* abs/2010.09470 (2020). arXiv: 2010.09470. URL: https://arxiv.org/abs/2010.09470.

[2] Stefan Axelsson. "The base-rate fallacy and the difficulty of intrusion detection". In: *ACM Transactions on Information and System Security* 3.3 (2000), pp. 186–205. DOI: 10.1145/357830.357849.

[3] Robert F Erbacher. "Base-rate fallacy redux and a deep dive review in cybersecurity". In: *arXiv preprint arXiv:2203.08801* (2022).