D5.3 Security Risk Assessment Algorithms

Work Package:	WP5		
Lead partner:	CERTH		
Author(s):	Gerasimos Antzoulatos (CERTH), Grigoris Stathopoulos (CERTH), Ilias Gialampoukidis (CERTH), Stefanos Vrochidis (CERTH)		
Due date:	31/05/2022		
Version number:	1.0	**Status:**	Final
Dissemination level:	Public		

Project Number:	883284	**Project Acronym:**	7SHIELD
Project Title:	Safety and Security Standards of Space Systems, ground Segments and Satellite data assets, via prevention, detection, response and mitigation of physical and cyber threats		
Start date:	September 1st, 2020		
Duration:	30 months		
Call identifier:	H2020-SU-INFRA-2019		
Topic:	SU-INFRA01-2018-2019-2020 Prevention, detection, response and mitigation of combined physical and cyber threats to critical infrastructure in Europe		
Instrument:	IA		

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 883284

Revision History

Revision	Date	Who	Description
0.1	18/04/2022	CERTH	First release of the template. ToC and assignments finalisation.
0.2	29/04/2022	CERTH	First draft of sections 3 and 4
0.3	06/05/2022	CERTH	First draft of sections 1 and 2
0.4	13/05/2022	CERTH	Second draft of sections 1 to 4
0.5	17/05/2022	CERTH	Executive summary and conclusions
0.6	19/05/2022	CERTH	Final proofreading. Fixed acronyms, references, captions, document styles.
0.7	20/05/2022	CERTH	Release version for internal review
1.0	31/05/2022	CERTH	Final release

Quality Control

Role	Date	Who	Approved/Comment
Internal review	30/05/2022	DEIMOS	Document accepted; only minor changes suggested
Internal review	30/05/2022	CeRICT	Document accepted; only minor changes suggested

Disclaimer

This document has been produced in the context of the 7SHIELD Project. The 7SHIELD project is part of the European Community's Horizon 2020 Program for research and development and is as such funded by the European Commission. All information in this document is provided 'as is' and no guarantee or warranty is given that the information is fit for any particular purpose. The user thereof uses the information at its sole risk and liability. For the avoidance of all doubts, the European Commission has no liability with respect to this document, which is merely representing the authors' view.

Executive Summary

This deliverable includes the outcomes of the activities carried out in the framework of the *Task 5.3 – Security Risk Assessment Algorithms for Decision Support* within WP5 of 7SHIELD project. The main objective of this task is to design and develop a module, called Crisis Classification (CRCL), that encompasses advanced machine learning processes so as to provide real-time assessments of the severity level of a crisis in the Critical Infrastructures (CIs) and, particularly, in the Satellite Ground Segments (SGS).

The Cyber-Physical systems (CPS) have been considered as a core of Cis, therefore, their protection from complex physical and cyber crisis events is a major challenge. The CRCL aims to dynamically assess the severity level of complex events and updates the Situational Picture (SP) of a particular critical ground segment, supporting crisis operators to decision-making processes.

Initially, an overview of the finalized user requirements is presented which, along with the key result objectives, are used as a guideline for the implemented solutions. Then, an overview of the machine learning practices for Risk Assessment in Critical Infrastructures is presented. A detailed presentation of the proposed framework that was developed and deployed follows. The framework relies on the utilisation of Machine Learning approaches that enable to estimate of the severity level of an attack (physical, cyber or combined). To achieve this, the preliminary step of the creation of adequate annotated datasets is required to train the models. Then, the interconnections between CRCL, the other 7SHIELD modules and the 7SHIELD platform as well as the benefits are presented. The evaluation experiments that were carried out throughout the development process and the produced results, both quantity and quality wise, are also reported here in order to explore the applicability of the final product. The performance of the machine learning algorithms has been evaluated in terms of their accuracy, F1 score, recall and precision.

Table of Contents

List of Figures

List of Tables

Definitions and acronyms

CI	Critical Infrastructure
CIP	Critical Infrastructure Protection
CEP	Complex Event Processing
C/P	Cyber/Physical
CRCL	Crisis Classification
C3	Command Control and Coordination
DBMS	Data Base Management System
DoA	Description of Action
DTC	Decision Tree Classifier
EC	European Commission
EU	European Union
GA	Grant Agreement
IDMEF	Intrusion Detection Message Exchange Format
JSON	JavaScript Object Notation
LR	Linear Regression
ML	Machine Learning
PC	Project Coordinator
RDBMS	Relational Data Base Management System
RF	Random Forest
SGS	Satellite Ground Station
SGSA	Satellite Ground Segment Asset
SVM	Support Vector Machine
UAF	Unified Alert Format
UI	User Interface
WP	Work Package
XGBoost	Extreme Gradient Boosting

1. Introduction

1.1. Scope of this deliverable

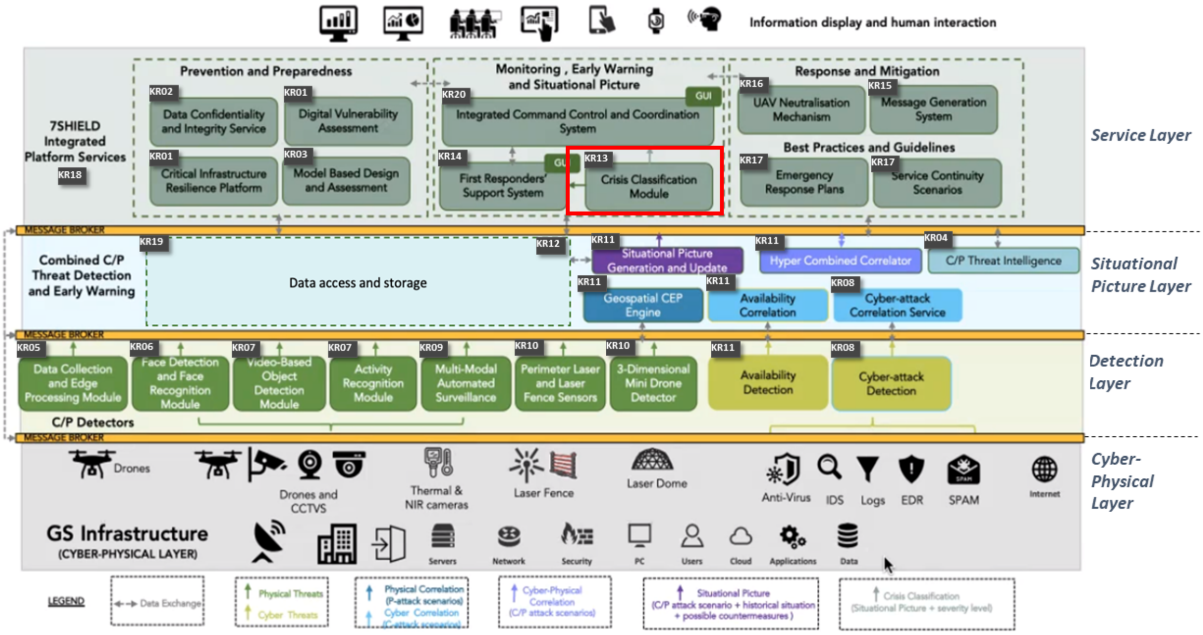

Information display and human interaction
7SHIELD Integrated Platform Services
KR18
Prevention and Preparedness
KR02
Data Confidentiality and Integrity Service
KR01
Digital Vulnerability Assessment
KR01
Critical Infrastructure Resilience Platform
KR03
Model Based Design and Assessment
Monitoring , Early Warning and Situational Picture
KR20
GUI
Integrated Command Control and Coordination System
KR14
GUI
First Responders' Support System
KR13
Crisis Classification Module
Response and Mitigation
KR16
UAV Neutralisation Mechanism
KR15
Message Generation System
Best Practices and Guidelines
KR17
Emergency Response Plans
KR17
Service Continuity Scenarios
Service Layer
MESSAGE BROKER
Combined C/P Threat Detection and Early Warning
KR19
Data access and storage
KR12
KR11
Situational Picture Generation and Update
KR11
Hyper Combined Correlator
KR04
C/P Threat Intelligence
KR11
Geospatial CEP Engine
KR11
Availability Correlation
KR08
Cyber-attack Correlation Service
Situational Picture Layer
MESSAGE BROKER
KR05
Data Collection and Edge Processing Module
KR06
Face Detection and Face Recognition Module
KR07
Video-Based Object Detection Module
KR07
Activity Recognition Module
KR09
Multi-Modal Automated Surveillance
KR10
Perimeter Laser and Laser Fence Sensors
KR10
3-Dimensional Mini Drone Detector
KR11
Availability Detection
KR08
Cyber-attack Detection
C/P Detectors
Detection Layer
MESSAGE BROKER
Drones
Drones and CCTVS
Thermal & NIR cameras
Laser Fence
Laser Dome
Anti-Virus
IDS
Logs
EDR
SPAM
Internet
GS Infrastructure
(CYBER-PHYSICAL LAYER)
Servers
Network
Security
PC
Users
Cloud
Applications
Data
Cyber-Physical Layer
LEGEND
Data Exchange
Physical Threats
Cyber Threats
Physical Correlation (P-attack scenarios)
Cyber Correlation (C-attack scenarios)
Cyber-Physical Correlation (C/P attack scenarios)
Situational Picture (C/P attack scenario + historical situation + possible countermeasures)
Crisis Classification (Situational Picture + severity level)

Figure 1-1 – High Level Logical Architecture of 7SHIELD

The main objective of 7SHIELD is to provide to the European Ground Segment facilities a holistic framework enabling to confront complex cyber and physical threats by covering all the macro stages of crisis management, namely pre-crisis, crisis and post-crises phases. To achieve this objective, a multi-layer architecture (Figure 1-1) has been developed which encompasses detection technologies to raise alerts from captured physical and/or cyber-attacks (*Detection Layer*) and correlation and fusion techniques that generate malicious events relied on those alerts (*Situational Picture Layer*). In this layer, the Situational Picture of the Ground Segments is generated and updated upon the crisis events. The severity level assessment of those crisis events is provided by the ***Crisis Classification*** (***CRCL***) module and further enhances dynamically (in real-time) the Situational Picture which is provided to the crisis managers and other stakeholders in the Command, Control and Coordination (C3) System. The Crisis Classification module is laid in the Service Layer of the 7SHIELD Logical architectural schema. The decision support services in this layer are responsible for prevention, management and mitigation activities used by the experts to tackle threats associated with cyber-physical attacks in the Satellite Ground Segment domain.

The scope of this deliverable is to focus on the detailed description of the Crisis Classification module in terms of its architecture, functionalities and interactions with other 7SHIELD modules. This deliverable covers the outcomes of the ***Task 5.3 Security Risk Assessment Algorithms for Decision Support*** which aims to design and develop a module enhancing the decision-making processes, by the real-time assess the severity of an ongoing physical and/or cyber-attack in critical satellite and ground segments. This task is mapped to the Key Result ***KR13: Crisis classification module***.

1.2. User requirements for risk assessment

Following the submission of the 1st and the 2nd round of the elicitation of the End User Requirements (URs) report the following requirements were examined regarding their relevance to KR13 in the context of the Task 5.3. Those are presented in Table 1-1.

ID	Type	Priority	Description	AC	KR
FR_SW_10	Functional	ESSENTIAL	7SHIELD must produce alerts based on crisis level classification and with indicator of crisis (e.g., red, orange, yellow colour).	AC_041	KR13, KR20
FR_SCE_10	Functional	OPTIONAL	7SHIELD could make decision of criticality through object recognition and operate autonomously in case of critical incident.	-	KR13
NFR_PERF_18	Non-Functional	ESSENTIAL	Escalation of incident (time for operator to create incident notification and alert competent authorities) must be completed in under 3 minutes.	AC_060	KR11, KR13, KR14
NFR_PERF_19	Non-Functional	ESSENTIAL	Percentage of alerts automatically linked to recommendations on minimizing impact propagation and supporting decisions must be at least 80%.	AC_061	KR11, KR13

Table 1-1 - User Requirements related to KR13

The essential functional requirement **FR_SW_10** states that "*7SHIELD must produce alerts based on crisis level classification and with indicator of crisis (e.g., red, orange, yellow colour)*" which is aligned with the main objective of the KR13 thus, it was recognized as the main UR which will be fully supported by KR13. Its description is straight-forward setting the ultimate goal of the CRCL module which is to provide real-time assessments of the severity of a crisis generated by a physical and/or cyber-attack in critical satellite and ground segments by fusing and analysing multimodal information and data.

The optional requirement **FR_SCE_10** states that "*7SHIELD could make decision of criticality through object recognition and operate autonomously in case of critical incident*" which is partially covered by KR13 in the sense that the CRCL module can provide assessments concerning the criticality of an ongoing event by the analysis of the detected and recognised objects. The Situational Picture can be updated automatically by CRCL and the Emergency Response Plans will be activated enabling the operators to respond and mitigate a critical incident.

The analysis of the received information from the tools of the Situational Picture layer is a straightforward process and longs some seconds. Additionally, every time where the Situational Picture is generated or updated, by including new malicious events, the CRCL assesses the severity level of the crisis in order to raise the awareness of the operators and crisis managers to activate the necessary emergency response plans. Hence, the non-functional essential requirements, namely the **NFR_PERF_18** and **NFR_PERF_19** will be covered.

1.3. Reference to other activities and documents

This deliverable D5.3 has direct or indirect interdependences with other WP and deliverables within the project:

- In the deliverables *D2.2 Consolidation of Stakeholder Requirements* and *D2.4 Use cases and requirements v2)*, released at M6 and M16 respectively, in which the definition and elicitation of the end-users' requirements of the 7SHIELD system have been carried out.
- In the deliverables *D4.2 - 7SHIELD Combined Physical and Cyber Threat detection* and *D6.2 – System integration and interoperability v1 (1st prototype)*, released at M9 and M10 respectively, the Unified Alert Format (UAF) messages and the messages to be exchanged between modules in the 7SHIELD ecosystem have been described. Part of the messages from the SPGU module is consumed and analysed by the CRCL module. Also, in those deliverables, the messages that will be published by CRCL have been described.
- The detections and correlations that are carried out from the modules of WP4 and affect the situational picture have direct interaction with the operation of the CRCL module.
- The provided solution within this deliverable will be implemented and integrated into the 7SHIELD platform in the context of the WP6 (System Integration). Furthermore, it will be tested and evaluated in the various Pilot Use Cases in the framework of WP7 (Pilot implementation, evaluation, and Training).
- The outcomes of the T5.3 will contribute, through scientific and research publications, to the dissemination and communication activities of the 7SHIELD project (WP8).

1.4. Deliverable structure

This deliverable consists of the following sections:

- In Section 1, the scope of the deliverable along with a mention to the user requirements and needs that this task (T5.3) should meet and references to other activities and documents are presented.
- In Section 2, a brief overview of the machine learning techniques applied in the Risk Assessment and the Disaster/Crisis Management is illustrated.
- Section 3 is dedicated to the presentation of the Crisis Classification module including its architecture, the design, development and evaluation processes. Additionally, the accessory tool, namely the Annotation Tool, which aims to create annotated datasets employing the user's experience and knowledge is thoroughly introduced.
- In Section 4 the experimental results and evaluations that have already carried out in order to estimate the performance of the Crisis Classification module are exhibited thoroughly.
- In Section 5, the main outcomes and conclusions of this work are drawn along with thoughts for future extensions of this module.

2. Overview of Risk Assessment in Critical Infrastructures

This section aims to describe the basic concepts behind the Risk Assessment processes in Critical Infrastructures (CI) relied on the employment of advanced Machine Learning approach. The rise of effective Machine Learning methodologies for the prevention, prediction and detection of extreme hazardous events generated by natural, or human-made (physical or cyber) sources, has started to contribute to the strengthening of our societies' resilience in various sectors and in the Critical Infrastructure Protection (CIP).

CIs can be designated as the physical structures, facilities, networks and other assets which provide services that are essential to the social and economic functioning of a community or society [1]. Under the umbrella term *Critical Infrastructure,* resources from various sectors are encompassed that are necessary for the operation of societies, such as energy capacities, information and communication technologies, utility services, water facilities, transport, health care, public administration premises and services, as well as services and infrastructures of the private sector [2]. Therefore, their functional continuity is required to ensure the security of a given nation, its economy, and the public’s health and/or safety. Conversely, their disruption in operation or destruction could cause long-term harmful consequences for the basic values of the society, cascading effects on other interdependent systems resulting in catastrophic results.

Nowadays, the crisis panorama has changed and diversify increasingly from “traditional” crises generated by natural hazards to technology-driven crises generated by cyber-attacks, or a combination of them ([3], [4]). The unexpectedly large scale of the extreme natural events in terms of their severity and frequency, the trans-boundary and cross-sectoral nature of new or unprecedented crises, compose a challenging and changing landscape in disaster and risk management [5]. In Global Assessment Report on Disaster Risk Reduction 2019, has been underlined the need to move beyond the conventional definition for the disaster risk, re-examine and re-assess the risk, by taking into consideration the pluralistic nature of it: in multiple dimensions, at multiple scales and with multiple impacts [6]. Furthermore, the advances of new technologies, from one side intensifies the potential threats and attacks and from the other, provides empowered solutions to address them and strengthen the resilience in human societies and CIs. Recent technological innovations like IoT, 5G, unmanned aircraft vehicles, and artificial intelligence have brought immense benefits and contributed further efficiencies to CI operations. However, they have posed serious threats facilitating the malicious actors interested in disrupting CI operations. Particularly, in the CIs which are becoming increasingly complex, automated, and interconnected, thereby new vulnerabilities have been introduced exposing them to malicious physical activities ([3], [4], [7]). Reducing the vulnerabilities and improving resilience of CIs have become a priority for the authorities around the globe in order to enhance the CIs protection and their operational continuity.

In CPS, physical security is very often disregarded as the main attention is focused on mitigation actions and security countermeasures oriented to response to cyber-attacks [8]. However, in the last decade, physical security has been evolved to be more challenging, compared to previous decades, as there are more sensitive data storages and devices available (like USB drives, laptops, smartphones, tablets, etc.) that are vulnerable to physical threats [9]. Therefore, physical security in CIs should go beyond controlling the access of authorized personnel to the premises and adopt into the security systems innovative surveillance solutions that provide recognition and monitoring of human activities inside and outside critical areas [8].

2.1. Machine Learning in Disaster Management Cycle

Descriptive Machine Learning methods focus on the Response and Recovery phases of the Disaster Management Cycle while the Predictive Machine Learning methods concentrate to provide forecasting assessments of a natural disaster, enhancing the preparedness and mitigation processes of the Disaster Management Cycle (Figure 2-1) [17].

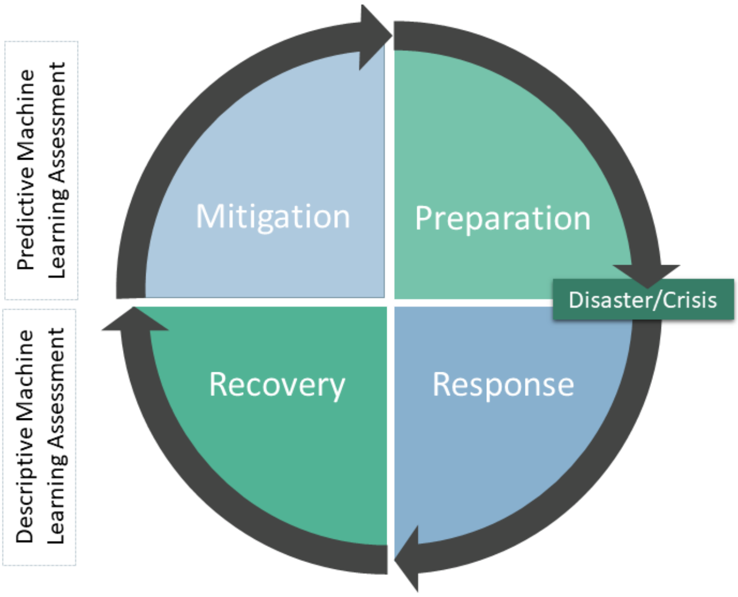

Predictive Machine Learning Assessment
Descriptive Machine Learning Assessment
Mitigation
Preparation
Disaster/Crisis
Recovery
Response

Figure 2-1 - Machine Learning approaches in relation to Disaster Management Cycle phases

Although the application of Machine Learning methodologies to tackle specific problem areas in disaster risk management dates back to a recent couple of decades, significant challenges still need to be addressed. Machine Learning methods have penetrated in a descriptive and/or predictive manner in all the phases of disaster/crisis management, contributing in various ways to the assessment of the hazard, exposure and vulnerability from natural and human-made disasters (([17], [18], [19], [20]).

Hence, one of the main challenges concerns the lack of required training data which limits the utilization of the machine learning algorithms to be trained in order for the latter to be able to predict or assess the risk of a crisis event. In the field of disaster/crisis management, the extreme events are rare, so the collection of reliable data during such events is often extremely difficult. Furthermore, the rapid increment in the amount of heterogeneous collected data does not necessarily imply that this problem will be repeal in the future ([11], [16]-[20]). As researchers pointed out more efforts for data collection will be required simultaneously with the establishment of data standardization protocols in order to enhance the collaboration, knowledge sharing and interoperability among different organizations, networked global data systems and stakeholders ([16]-[20]).

Getting motivated by this gap, the proposed annotation tool (Section 3.1) aims to involve the experts in the Satellite ground segments domain, by mapping their experience and knowledge into the characterization of hypothetical extreme physical (natural or human-made) events and cyber-attacks in terms of their severity and impact to the CI.

In Emergency situations, the estimation of a severity level of an involving crisis timely and seamlessly is crucial for the operators to efficiently response and take the appropriate countermeasures. Hence, recently, the utilisation of powerful machine learning techniques for strengthen the trackability and monitoring extreme natural hazardous events have been applied. For example, in [12] a multi-Layer Fusion framework, for Real-Time Fire Severity Assessment, based on knowledge extracted from the analysis of Fire Multimedia Incidents has presented. Recently, the analysis of remote sensing data from satellite images and GIS based data along with the utilisation of advanced machine learning algorithms can contribute to the flood mapping and monitoring of the flood hazard and risk in a specific region of interest [15]. Additionally, in [13] authors proposed an open-source holistic framework encompasses technological achievements that enables first responders and authorities to manage efficiently the pre-emergency phases of a hazardous natural event focusing on floods, wildfires and heatwaves relied on rule-based approach. In [14] an extension of it has been proposed where a unified multi-layer framework that encapsulates machine learning techniques in the risk assessment process. The aim was the analysis and fusion of dynamically heterogeneous information obtained from the field by covering pre-emergency and emergency phases of a crisis.

Relied on the aforementioned approach, in [16] is proposed a framework that enables to identify potential human-made threats, generated by using physical means. These physical malicious activities can be detected by heterogeneous sources (CCTVs, UAVs, etc.) and the live streaming analysis based on the advanced machine learning techniques can provide useful information to early warning crisis managers and operators to respond and mitigate the potential attack. In Section 3 a detailed description of this framework is provided.

3. 7SHIELD Security Risk Assessment Algorithms

In this section, the proposed framework which estimates the level of severity during physical threats ([16]) is described and extended in order to cover all the aspects of the 7SHIELD project in terms of the potential threats, namely physical, cyber and complex events. Firstly, a description of the adequate Annotation Tool is presented.

3.1. Annotation Tool

3.1.1. Architectural Schema

The Annotation Tool is a web application that its main purpose is to enable the creation of a machine learning training dataset, representing hypothetical attack scenarios, the potential consequences and the likelihood of them to take place in the 7SHIELD premises. Since the 7SHIELD pilot sites are unique, there are no public annotated datasets that are able to represent the current problem, hence, we decided to create custom ones, tailored to fit our needs.

Another focus for the Annotation Tool was to be as easy and fast to use as possible. The larger the annotated dataset, the best the results after the training procedure of the machine learning modules would be. A web application can be used in almost any device, as long as it has access to a web browser and an active network connection, thus there is no need for installation. By ensuring ease of accessibility and a clean, easy to use UI, the end user has the ability to annotate larger quantities of hypothetical scenarios, faster and more efficient. The Annotation Tool consists of several parts as illustrated in the following figure (Figure 3-1). The Scenario Builder is responsible for the creation of random generated attack scenarios. Those scenarios are then stored into MongoDB, an online database. They can then be retrieved by the Scenario Evaluator, where the user can annotate them using the web application UI. The annotated scenario is then stored again in MongoDB.

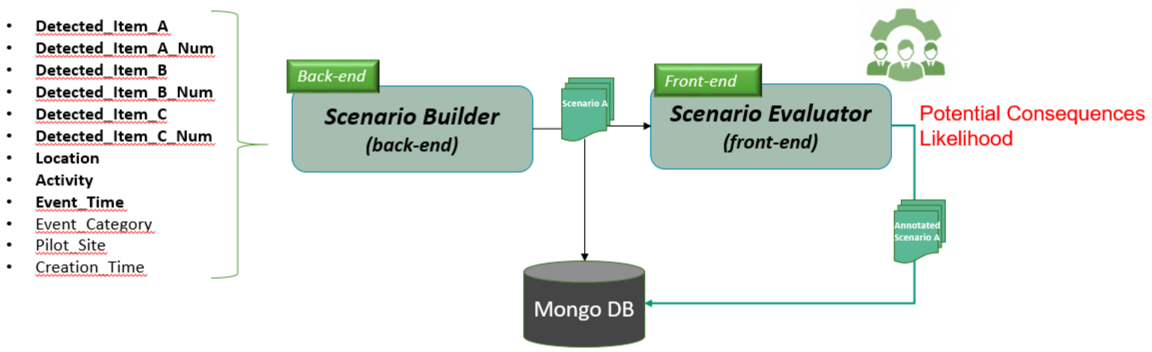
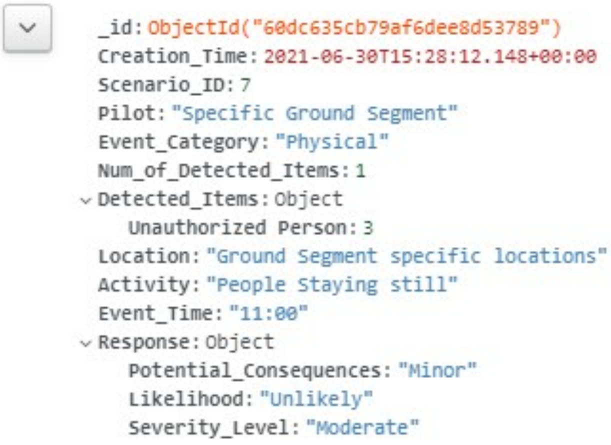


Figure 3-1 – A High Level Architecture of the Annotation Tool

3.1.2. Back-End Process

In order to create the random hypothetical scenarios, Annotation Tool uses a Python script that is able to mix and match hand-picked parameters, specific for each pilot site and attack type. When each type of information necessary for the definition of the scenario is generated, it is stored in a JSON-like format in MongoDB. Each scenario has the fields that are needed to describe the event taking place, such as the location and if some people or objects were detected. It also has some empty fields that are intended to get the information the user provides, after the characterization of the scenario takes place, such as the likelihood and the potential consequences.



```
  _id: ObjectId("60dc635cb79af6dee8d53789")
  Creation_Time: 2021-06-30T15:28:12.148+00:00
  Scenario_ID: 7
  Pilot: "Specific Ground Segment"
  Event_Category: "Physical"
  Num_of_Detected_Items: 1
⌄ Detected_Items: Object
     Unauthorized Person: 3
  Location: "Ground Segment specific locations"
  Activity: "People Staying still"
  Event_Time: "11:00"
⌄ Response: Object
     Potential_Consequences: "Minor"
     Likelihood: "Unlikely"
     Severity_Level: "Moderate"
```

Figure 3-2 - Example of an annotated event stored into MongoDB

Annotation Tool also has a build in Login system that enables the monitoring of the users that have access and can characterize the hypothetical events. In order for someone to log in and use the Tool, correct input of the credentials (username, password) is required.

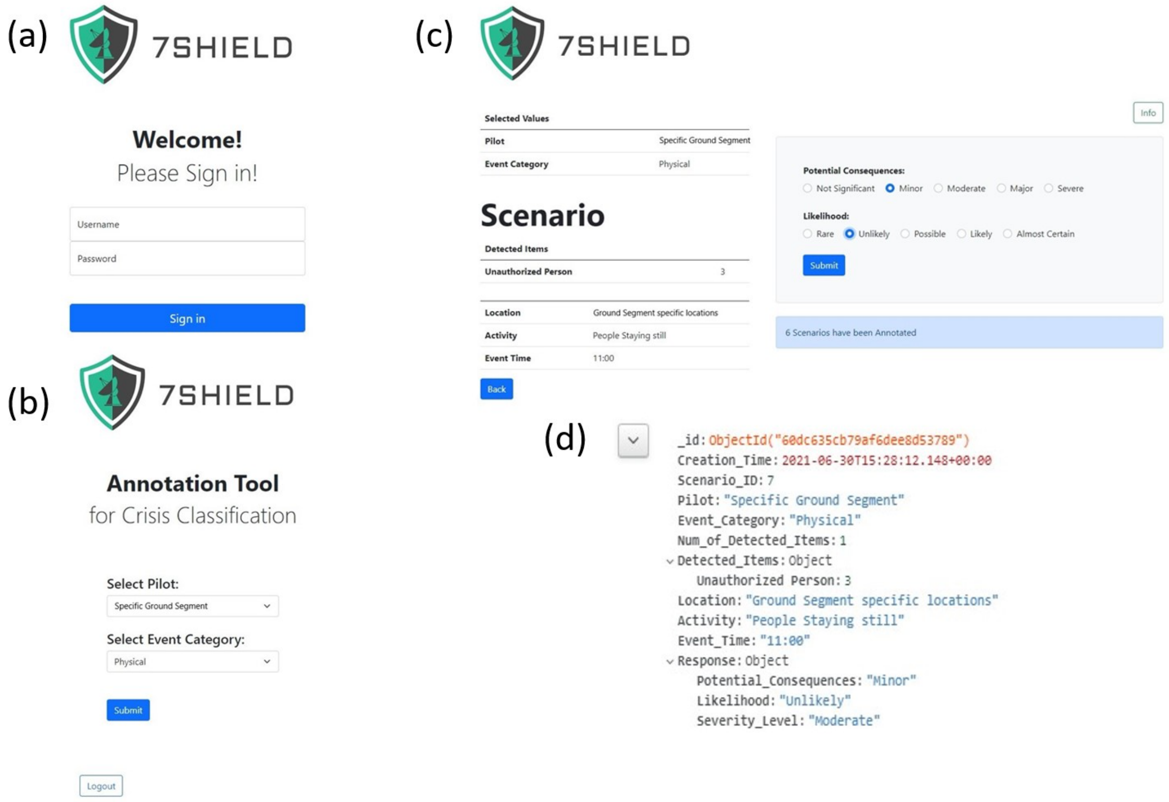
The main framework used to create this web application is Flask[1]. Flask is using python and is implemented on Werkzeug[2] and Jinja[3]. The advantages of using Flask web framework are many, including:

- Built-in development server and a fast debugger provided
- Lightweight
- Secure cookies are supported
- Templating using Jinja2
- Request dispatching using REST
- Support for unit testing built-in

[1] https://palletsprojects.com/p/flask/
[2] https://palletsprojects.com/p/werkzeug/
[3] https://palletsprojects.com/p/jinja/

3.1.3. Front-End User Interfaces

The front-end of the Annotation tool was built using Python, HTML, Bootstrap and Typescript. Bootstrap is a front-end open-source toolkit, featuring a collection of HTML, CSS and JavaScript tools for creating and building web pages and web applications. Typescript offers all of JavaScript's features and an additional layer, Typescript's type system, lowering the chance of bugs.



Select Pilot: Specific Ground Segment

Select Event Category: Physical

Selected Values	
Pilot	Specific Ground Segment
Event Category	Physical

Scenario

Detected Items	
Unauthorized Person	3

Location	Ground Segment specific locations
Activity	People Staying still
Event Time	11:00

Potential Consequences: ○ Not Significant ● Minor ○ Moderate ○ Major ○ Severe

Likelihood: ○ Rare ● Unlikely ○ Possible ○ Likely ○ Almost Certain

6 Scenarios have been Annotated

```
_id: ObjectId("60dc635cb79af6dee8d53789")
Creation_Time: 2021-06-30T15:28:12.148+00:00
Scenario_ID: 7
Pilot: "Specific Ground Segment"
Event_Category: "Physical"
Num_of_Detected_Items: 1
Detected_Items: Object
   Unauthorized Person: 3
Location: "Ground Segment specific locations"
Activity: "People Staying still"
Event_Time: "11:00"
Response: Object
   Potential_Consequences: "Minor"
   Likelihood: "Unlikely"
   Severity_Level: "Moderate"
```

Figure 3-3 - Overview of the Annotation Tool front and back end

A web application is a collection of static and dynamic web pages. A static web page is one that does not change when a site visitor requests it: The web server sends the page to the requesting web browser without modifying it. A dynamic web page on the other hand is modified by the server before it is sent to the requesting browser. The changing nature of the page is why it's called dynamic. Annotation Tool's front-end consists of the following web pages:

a) **Sign in page**: A welcome page where the end-user can input the credentials ("Username" and "Password") required in order to get access to the main interface (Figure 3-3 (a)). Wrong or missing credentials will result in an error popup message, since the user could not get authenticated.

b) **Pilot/Event category selection page:** A page where the end-user can use drop-down menus in order to select the pilot site and the event category of the hypothetical scenarios that will be generated for annotation (Figure 3-3 (b)). According to the Pilot Use Case scenarios, the available choices are the following:

Pilot	Event Category
FMI	Physical
SPACEAPPS	Cyber
SERCO	Cyber
DEIMOS	Physical or Cyber
NOA	Physical or Cyber

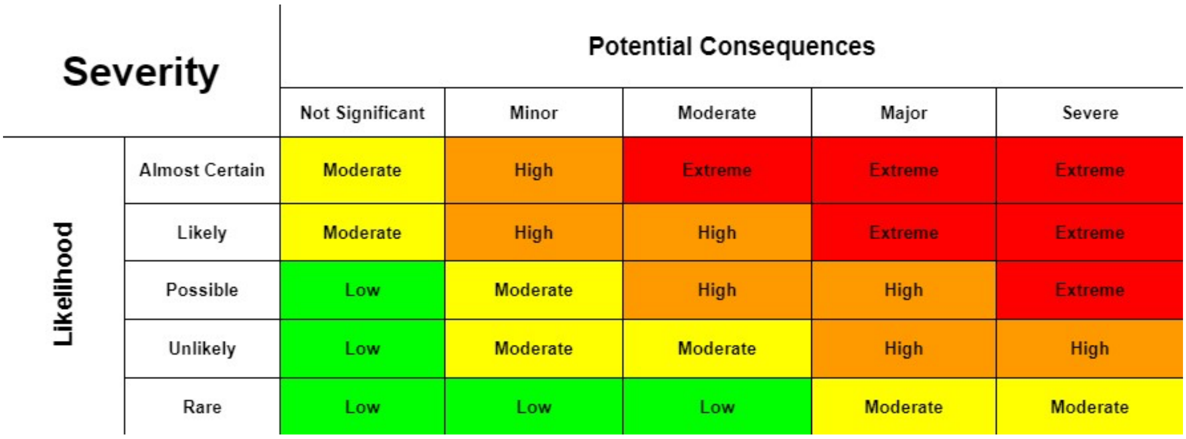
Table 3-1 – Event category per pilot use case

c) **Main interface – Annotation page:** This is the page where the random generated scenarios can be characterized by the end-users (Figure 3-3 (c)). On the left side of the page, under the "Scenario" tag there are values that represent the hypothetical attack scenario. Those values consist of:

1. "Detected Items", where all the possible threats (e.g. Unauthorized Person, Car, Motorcycle, etc.) and their population are listed
2. "Location", where the location of the "Items" detected or the asset that is targeted is declared
3. "Activity", for the activity of the person(s) detected
4. "Event Time" specifies the time when the event is taking place

On the top right of the page, a popup "info" button is present, providing the user useful information on how to use the Tool. Bellow it, a multiple selection box enables the user to characterize the scenario by choosing one out of five possible values for each one of the fields "Potential Consequences" and "Likelihood". The two later choices are used for the calculation of the overall "Severity Level" of the event.

In order to get a final JSON file with the annotated scenario like the one in (Figure 3-3 (d)), "Severity Level" must be calculated. This is an automated back-end procedure, where a custom risk matrix is used, adjusted to the needs of 7SHIELD.



Severity		Potential Consequences				
		Not Significant	Minor	Moderate	Major	Severe
Likelihood	Almost Certain	Moderate	High	Extreme	Extreme	Extreme
	Likely	Moderate	High	High	Extreme	Extreme
	Possible	Low	Moderate	High	High	Extreme
	Unlikely	Low	Moderate	Moderate	High	High
	Rare	Low	Low	Low	Moderate	Moderate

Figure 3-4 - Risk Matrix used to calculate severity level

3.2. Crisis Classification Module

3.2.1. High Level Architectural Schema

The accurate and timely estimation of the severity of the crisis is an ultimate goal for authorities to effectively respond and handle an ongoing crisis. In the 7SHIELD ecosystem the aim of the Crisis Classification module is to encompass methodologies for multi-level crisis assessment relying on multimodal information and data fusion. Particularly, it will employ existing machine learning approaches, modified specifically for the needs of 7SHIELD, in order to assess and classify the severity of ongoing crisis events.

The main focus of the design and development of Crisis Classification module is to enhance the decision-making processes, by real-time assessing the severity level of an ongoing physical and/or cyber-attack in critical space systems, ground segments and satellite data assets. The module will incorporate multi-level fusion techniques so as to analyse multiple types of data, classify crisis events utilising machine learning techniques and extend the decision support processes of the responders. Severity level will be estimated by combining the available sources of data and the outcomes of the detection technologies for physical and cyber threats.

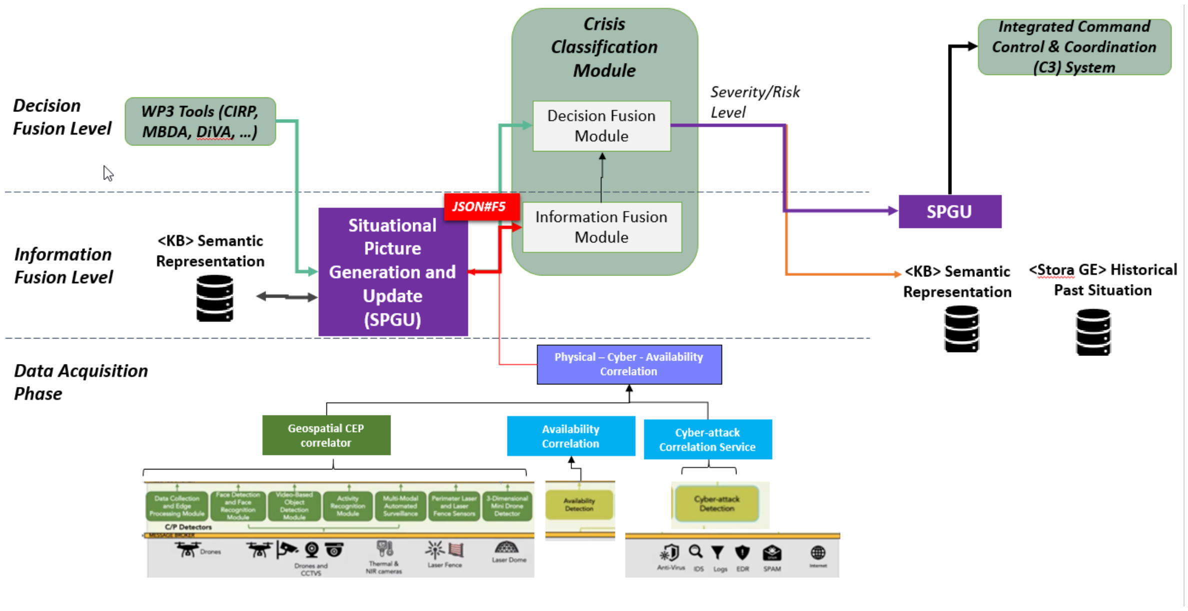

Crisis Classification Module
Decision Fusion Level
WP3 Tools (CIRP, MBDA, DiVA, …)
Decision Fusion Module
Severity/Risk Level
Integrated Command Control & Coordination (C3) System
JSON#F5
Information Fusion Module
SPGU
Information Fusion Level
<KB> Semantic Representation
Situational Picture Generation and Update (SPGU)
<KB> Semantic Representation
<Stora GE> Historical Past Situation
Data Acquisition Phase
Physical – Cyber - Availability Correlation
Geospatial CEP correlator
Availability Correlation
Cyber-attack Correlation Service
Data Collection and Edge Processing Module
Face Detection and Face Recognition Module
Video-Based Object Detection Module
Activity Recognition Module
Multi-Modal Automated Surveillance
Perimeter Laser and Laser Fence Sensors
3-Dimensional Mini Drone Detector
C/P Detectors
MESSAGE BROKER
Availability Detection
Cyber-attack Detection
Drones
Drones and CCTVS
Thermal & NIR cameras
Laser Fence
Laser Dome
Anti-Virus
IDS
Logs
EDR
SPAM
Internet

Figure 3-5 - High level architectural schema of Crisis Classification

The Crisis Classification module consists of two main components, namely the *Information Fusion Module* and the *Decision Fusion Module* (Figure 3-5). The former consumes information from the Situational Picture of a Ground Segment regarding the real-time conditions for a malicious event (critical). This component estimates the severity level of the crisis event that is in progress by using machine learning techniques and proceeds it to the Decision Fusion module. The goal of this component is to enhance the severity level by combining information from the assets, their criticality and vulnerabilities in a rule-based approach. The final estimation of the severity/risk level of a particular crisis event updates the Situational Picture by sending the appropriate message to the SPGU module. Also, the Knowledge Base and the database which stores the historical Situational Picture events will be updated accordingly as it is illustrated in the above figure.

It should be noted that the alerts that are generated by the detectors, physical and cyber ones, in the Detection layer, are correlated via physical, cyber and combined correlators to produce more sophisticated events. Then, these correlated events are able to generate or update the Situational Picture concerning the status of the Ground Segment due to malicious or abnormal events (Figure 3-5). Hence, the Crisis Classification module will be triggered automatically every time that a new situation emerges by SPGU module. The severity level will be updated and published to SPGU. The final receiver of this process is the Command, Control and Coordination (C3) user interface where the information from the severity level will be visualised and raise awareness among the operators.

3.2.2. Information Fusion Module

The Information Fusion module consists of the 1st step of the analysis inside the CRCL module. It relies on the utilisation of various Machine Learning approaches for risk assessment. The aim is to fuse information from various detected items and events that have been identified by detectors and correlated by the 7SHIELD correlators in order to assess the severity level of the destructive events. The exploitation of the capabilities of Machine Learning approaches to "learn" from historical data and fit the behaviour of the models according to the current situation permits the develop of a module that will be reliable and robust to assess the severity level of a crisis in real-time.

The training of the ML models can be performed by the usage of an annotated dataset. In our case, the target variable is the severity levels (Low, Moderate, High, Extreme) and the classification of each event in those categories can be done based on pre-defined features extracted from the events. After the training process, each ML model will be evaluated over its performance to classify the "unknown" data. In our case, new crisis events appear and the trained models classify them in terms of their severity. In the end, the best performance ml model will be chosen to classify new physical or cyber events. In the following sections the aforementioned processes will be described in more details.

3.2.2.1. Training process

Training a machine learning (ML) model is a process in which a machine learning algorithm is fed with training data from which it can learn. That is the primary step in machine learning and results in a model that can be then validated, tested and deployed (Figure 3-6). The performance of the model during training will determine how well it will work when it is eventually put into an application for the end-users.

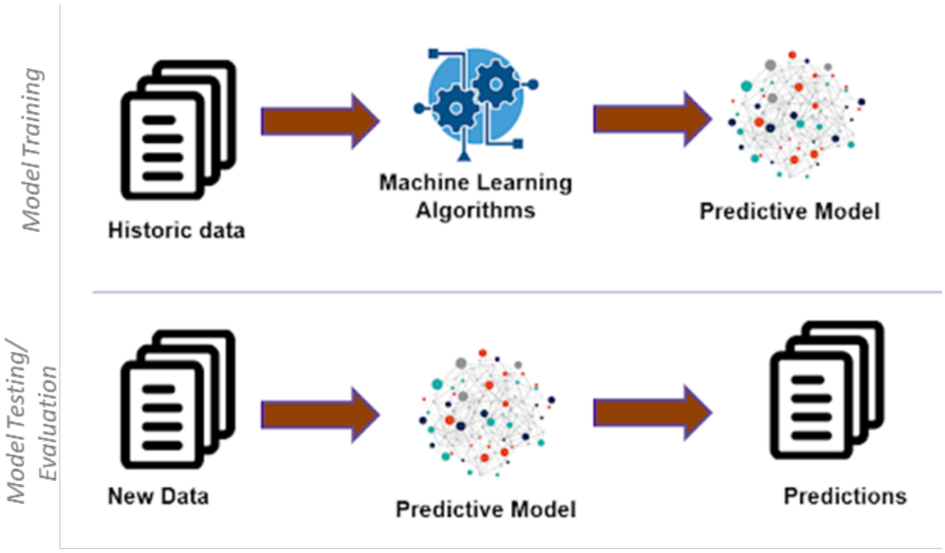


Figure 3-6 - General process of training and testing a ML algorithm [11]

Due to the nature of the problem that 7SHIELD presents, the uniqueness of each space ground segment and the specific requirements, there is no suitable open-source dataset that can efficiently be used to train the machine learning algorithms of the Crisis Classification module. In order to overcome that problem, we used Annotation Tool to effectively create our own annotated dataset. With the help of experts in the safety of each ground segment, we managed to collect a total of 1088 cyber and 762 physical hypothetical attack scenarios annotated.

The training process of the Information Fusion module includes a total of five different machine learning methodologies. Those are:

- Linear Regression (Ridge Classifier)

  One of the fundamental supervised machine-learning algorithms due to its relative simplicity and well-known properties. Ridge Regression is a method of estimating the coefficients of multiple-regression models in scenarios where linearly independent variables are highly correlated.

- Support-Vector Machine (SVM)

  Supervised learning model with associated learning algorithms that analyze data for classification and regression analysis. SVMs are one of the most robust prediction methods.

- Random Forest Classifier

  Classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of the dataset.

- Decision Tree Classifier

  Amongst the most popular machine learning algorithms given their intelligibility and simplicity. Classification trees are tree models where the target variable can take a discrete set of values. In these trees, leaves represent class labels and branches represent conjunctions of features that lead to those class labels.

- XGBoost Classifier (eXtreme Gradient Boosting)

  Open-source software library which provides a regularizing gradient boosting framework. It offers amongst other a clever penalization of trees and a proportional shrinking of leaf nodes, and it has gained much popularity recently as the choice algorithm of many winning teams of machine learning competitions.

In order to achieve the best possible trained model with the limited data sample created by the Annotation Tool (Section 3.1), cross-validation was used during the training of each method. The general procedure of a k-fold cross-validation is as follows:

1. Shuffle the dataset randomly
2. Split the dataset into k groups
3. For each group:
    - Take the group as a hold out or test data set
    - Take the remaining groups as a training data set

- Fit a model on the training data set and evaluate it on the test set
- Retain the evaluation score and discard the model

4. Summarize the skill of the model using the sample of model evaluation scores

Each observation in the data sample is assigned to an individual group and stays in that group for the duration of the procedure. This means that each sample is given the opportunity to be used in the hold out set 1 time and used to train the model k-1 times.

Hyperparameter optimization or tuning is another way to ensure that the model can optimally solve the machine learning problem. The same kind of machine learning model can require different constrains, weights or learning rates to generalize different data patterns. Hyperparameter optimization finds a tuple of hyperparameters that yields an optimal model which minimizes a predefined loss function on given independent data. The objective function takes a tuple of hyperparameters and returns the associated loss. To estimate this generalization performance, cross-validation was used.

After the training process with the annotated dataset, Crisis Classification is able to identify and save the best performing model for a specific Space Ground Segment, choosing the most optimal hyperparameters. To compare the best versions of each algorithm, predictions are made on the test data and results such as F1-score and accuracy are taken into account.

3.2.2.2. Experimental evaluation process

As mentioned in the paragraph above, Crisis Classification module needs to evaluate the training results. After the cross-validation of each different machine learning algorithm and the Hyperparameter tuning, we need to select the best fitting method and the best version of that particular algorithm.

When performing classification predictions, there are four types of outcomes that could occur:

- **True positives (TP)** are when you predict an observation belongs to a class and it actually does belong to that class.
- **True negatives (TN)** are when you predict an observation does not belong to a class and it actually does not belong to that class.
- **False positives (FP)** occur when you predict an observation belongs to a class when in reality it does not.
- **False negatives (FN)** occur when you predict an observation does not belong to a class when in fact it does.

These four outcomes are plotted on a confusion matrix. A paradigm can be seen bellow (Figure 3-7).

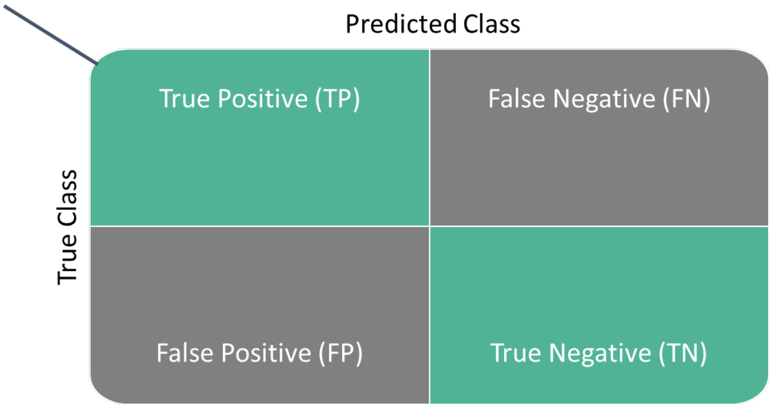


Figure 3-7 - Confusion Matrix example

The tree main metrics used to evaluate a classification model are **accuracy**, **precision** and **recall**. We also calculate and keep track of a 4th metric, **F1 score**[4].

- **Accuracy** is defined as the percentage of correct predictions for the test data. It can be calculated easily by dividing the number of correct predictions by the number of total predictions.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

- **Precision** is defined as the fraction of relevant examples (true positives) among all of the examples which were predicted to belong in a certain class.

$$Precision = \frac{TP}{TP + FP}$$

- **Recall** is defined as the fraction of examples which were predicted to belong to a class with respect to all of the examples that truly belong in the class.

$$Recall = \frac{TP}{TP + FN}$$

- **F1 Score** is defined as the harmonic mean of **Precision** and **Recall**. In order to calculate F1 Score we compute the average of precision and recall. Since it is the average, it means that it gives equal weight to Precision and Recall.

$$F1 - score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Through these metrics we are able to compare the performance of each machine learning algorithm, with each different set of parameters, for each pilot site and attack category (Cyber, Physical). In the end, the trained model with the best performance is serialized with Pickle and is saved in order to be used for severity prediction when an event is taking place in real time.

[4] https://www.sciencedirect.com/topics/engineering/confusion-matrix

Pickle is a module that implements binary protocols for serializing and de-serializing a Python object structure. When "pickling" a Python object hierarchy, it is converted into a byte stream. The inverse operation, "unpickling", is when the byte stream is converted back into an object hierarchy.

3.2.3. Decision Fusion Module

The Decision Fusion module is responsible to fuse the results of the Information Fusion module along with information originated from other modules of the Situational Picture layer or Service layer. Specifically, the assessments of the severity level for evolving crisis events will be enriched with information related to the vulnerabilities of the assets that are exposed to those attacks. For example, the data centre of a SGS is a critical area, hence the presence of an intruder there will be an extreme severity event. Hence, this module compiles specific user-defined rules in order to enrich the decisions generated by the Information Fusion module and increase the final severity level.

Furthermore, in the case that complex physical and cyber-attacks take place the CRCL should be able to classify those hybrid events in terms of their severity levels. As those malicious events have been correlated and encompassed in the same situational picture, thus the CRCL should provide the updated severity level for the whole situation. On the other hand, the Information Fusion module handles physical or cyber events separately, by producing different classifications of the severity level for each event.

Moreover, in the framework of the 7SHIELD project, the proper operation and availability of the sensors and hardware devices are monitoring via the Availability Detection Monitoring (ADM) module. The malfunction of them should be detected and early notified and has an impact to the overall estimation of the situational picture. Therefore, in the CRCL module the following empirical rule is utilised in order to assess the severity level caused by those events:

$$Severity\ level\ (Av) = \begin{cases} Low,\ \ if\ Availability\ Status = Up \\ Moderate,\ \ if\ Availability\ Status = Down \end{cases}$$

Hence, the Decision Fusion module should be capable of ensemble these decisions for severity level and provide a unified one, the overall Severity Level (Figure 3-8). A simple way to implement this process is to apply the rule of the maximum value:

$$Severity\ Level = max\{Severity\ level\ (Phy),\quad Severity\ level\ (Cy),\qquad Severity\ level\ (Av)\}$$

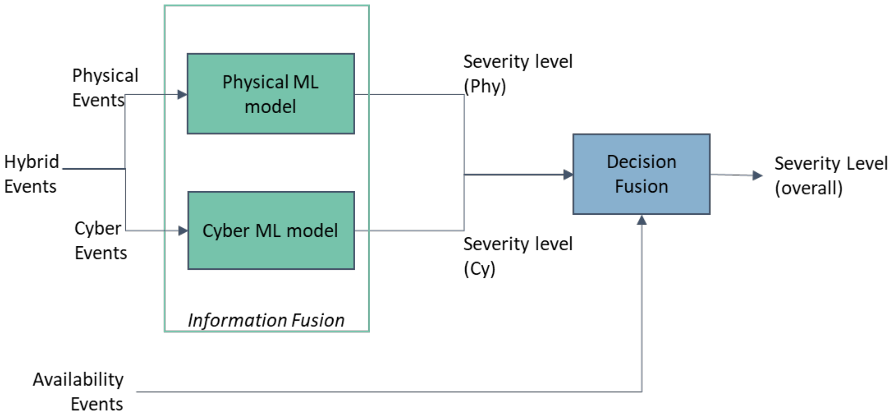


Figure 3-8 – Interaction between Information Fusion and Decision Fusion modules

3.2.4. Deployment and Interaction with Other Modules

Crisis Classification (CRCL) is connected with the rest of 7SHIELD modules through Apache Kafka. Kafka is a distributed event store and stream-processing platform, and aims to provide a unified, high-throughput, low-latency handling of real-time data feeds.

7SHIELD uses physical and cyber correlators in order to compare and group all the information that the detectors are providing when an event is taking place. Crisis Classification gets all the correlated data in real time through the Situational Picture Generation and Update module (SPGU). Using that input, CRCL calculates the severity level of the current situation, and it produces an output, in order for the Integrated C3 System to be updated.

4. Experimental Evaluations

4.1. End-User's Feedback for the Annotation Process

With the help of the Annotation Tool web application mentioned above, the experts for each space ground segment managed to annotate a variety of cyber and physical events. Those events were randomly generated hypothetical attack scenarios that were created through the back-end layer of the Annotation Tool. A total of 1088 cyber events and 762 physical events were annotated by the end users. In order to evaluate the annotated data, we generated multiple graphs that quantify the quality and the range of each dataset.

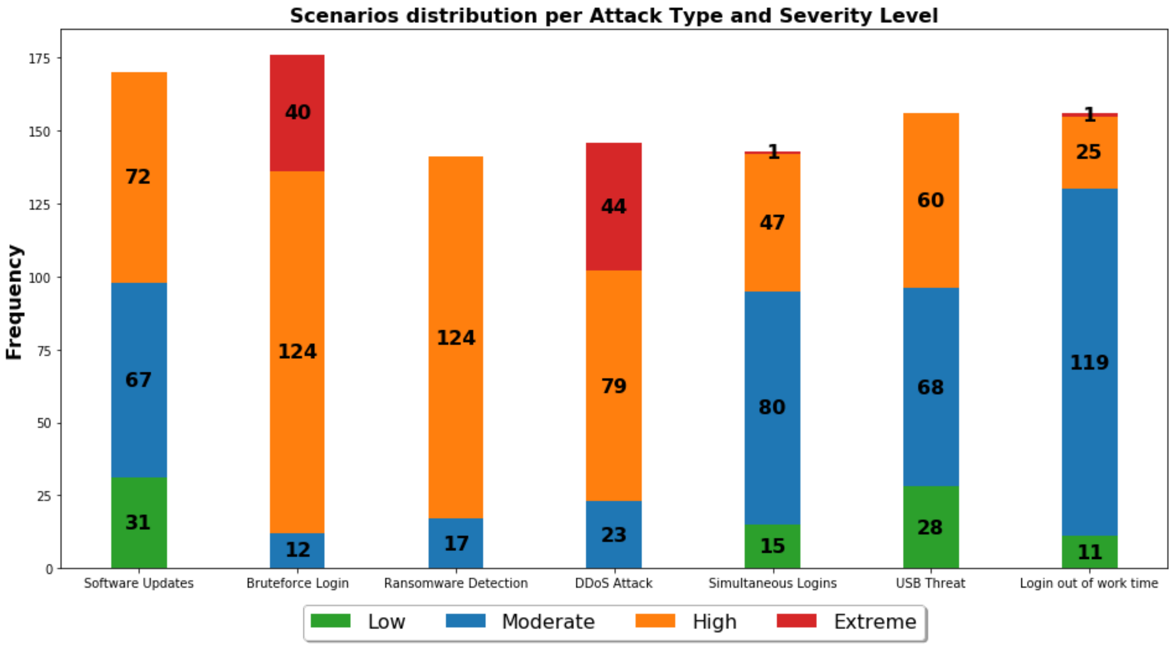


Figure 4-1 - Cyber scenarios distribution per attack type and severity level

In the above graph the main type of each cyber event that was annotated can be seen, along with the severity level that was assigned to that event by the end users. Some types of cyber-attacks or events lean towards the lower or the higher end of the severity level, as expected. For example, a lot of DDoS (Denial of Service) attacks were characterized with high or extreme severity where the "log in out of work time" type of event is less severe overall. Of course, there were also other factors involved as the time that the event was taking place and the asset it was targeted. Below, we can see the same type of graph generated for the physical events.

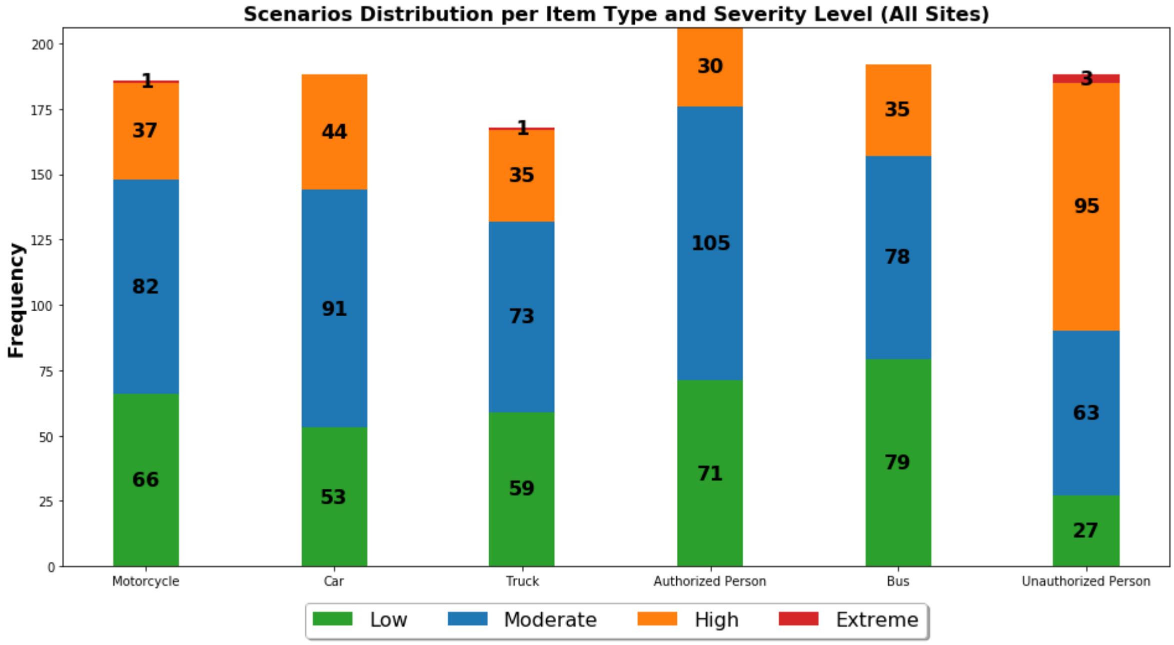


Figure 4-2 - Physical scenarios distribution per detected item type and severity level

In this graph, the physical items that can be detected and identified by the 7SHIELD detectors can be seen, as well as the frequency each one was present in the random generated events. The severity level of the scenarios that involved one or more of those items is also assigned to each category.

For each space ground segment, according to the needs of the project, either physical, cyber or both types of events were annotated. The total scenarios that were characterized for each pilot site according to the event type are:

- NOA (281 physical, 230 cyber)
- DEIMOS (301 physical, 458 cyber)
- FMI (202 physical)
- SPACEAPPS (200 cyber)
- SERCO (200 cyber)

4.2. Crisis Classification Experimental Results

Using the annotated datasets for each pilot site it was possible to train and test various machine learning models. The cross-validation method k-fold was deployed in order to estimate the skill of each machine learning model, as well as the best set of parameters. It should be noted that a quite exhausted fine-tuning of the parameters for each machine learning technique had been carried out.Bellow, the results for each one of the 5 different machine learning models over the best set of parameters when trained and tested with physical and cyber annotated datasets in a particular pilot site will be exhibited. Each model was measured in terms of Precision, Recall, F1 Score and Accuracy. The parameters that can be seen for each model are the set with the highest Accuracy for that specific machine learning methodology.

To better evaluate the results that the machine learning methods produce when applied to each pilot site, confusion matrixes are generated during the testing phase of the training session. The events used for testing are 20% of the total events that were used for the attack scenarios. The rest 80% was used for the training procedure.

4.2.1. Experimental Results Over the NOA Pilot

4.2.1.1. Physical attack scenarios

The annotated dataset generated by 281 physical attack hypothetical scenarios divided into 80% (224) events that utilised for the classifiers training and the rest portion of them, 20% (57) events for testing and evaluation purposes. In the following table (Table 4-1) the experimental results over the testing dataset per classifier are illustrated. The results that correspond to each classifier came out by using the best set of parameters that the particular classifier succeed in terms of Accuracy.

Classifier	Best set of parameters	Precision	Recall	F1 Score	Accuracy
LR (Ridge)	(alpha=1, normalize=True, tol=0.001, solver="auto", random_state=42)	71.87%	78.85%	70.92%	78.84%
SVM	(kernel="linear", C=0.1, gamma=1, random_state=42)	74.04%	80.77%	**75.87%**	**80.76%**
RF	(criterion="gini", max_depth=110, max_features="auto", min_samples_leaf=10, min_samples_split=10, n_estimators=100, random_state=42, bootstrap=True)	60.92%	65.38%	62.92%	65.38%
DT	(criterion="gini", max_depth=1, max_leaf_nodes=2, min_samples_split=2, random_state=42)	64.69%	67.31%	65.23%	67.31%
XGBoost	(colsample_bytree=0.7, learning_rate=0.01, max_depth=2, min_child_weight=10, n_estimators=80, nthread=4, subsample=0.8)	66.11%	71.15%	68.11%	71.15%

Table 4-1– Performance of the ML Classifiers over NOA's physical dataset

Best performing algorithm for NOA pilot's physical dataset is SVM with linear kernel, which its accuracy is approximately around 80.76%. Linear Regression classifier exhibits the 2nd best performance among the classifiers in terms of the Accuracy and F1-score. Decision Trees and Random Forest exhibit quite poor performance in terms of the Accuracy (67.31% and 65.38% correspondingly) and F1-score (65.23% and 62.92% correspondingly).

The confusion matrix for the best performing algorithm (SVM) in the NOA pilot's physical dataset is illustrated in the Figure 4-3. The SVM classifier managed to classify correctly the majority of the hypothetical scenarios that present moderate severity levels. However, it fails to identify correctly the scenarios with the low severity levels. One potential explanation could be that the scenarios of those categories may be quite similar having very few differences, so it is difficult to classify them correctly even by humans. This conclusion is reinforced by observing the results of the other classifiers, at confusion matrices in Figure 4-4.

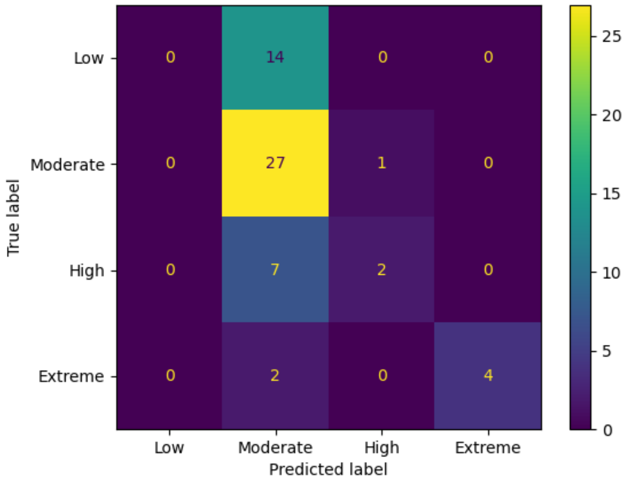
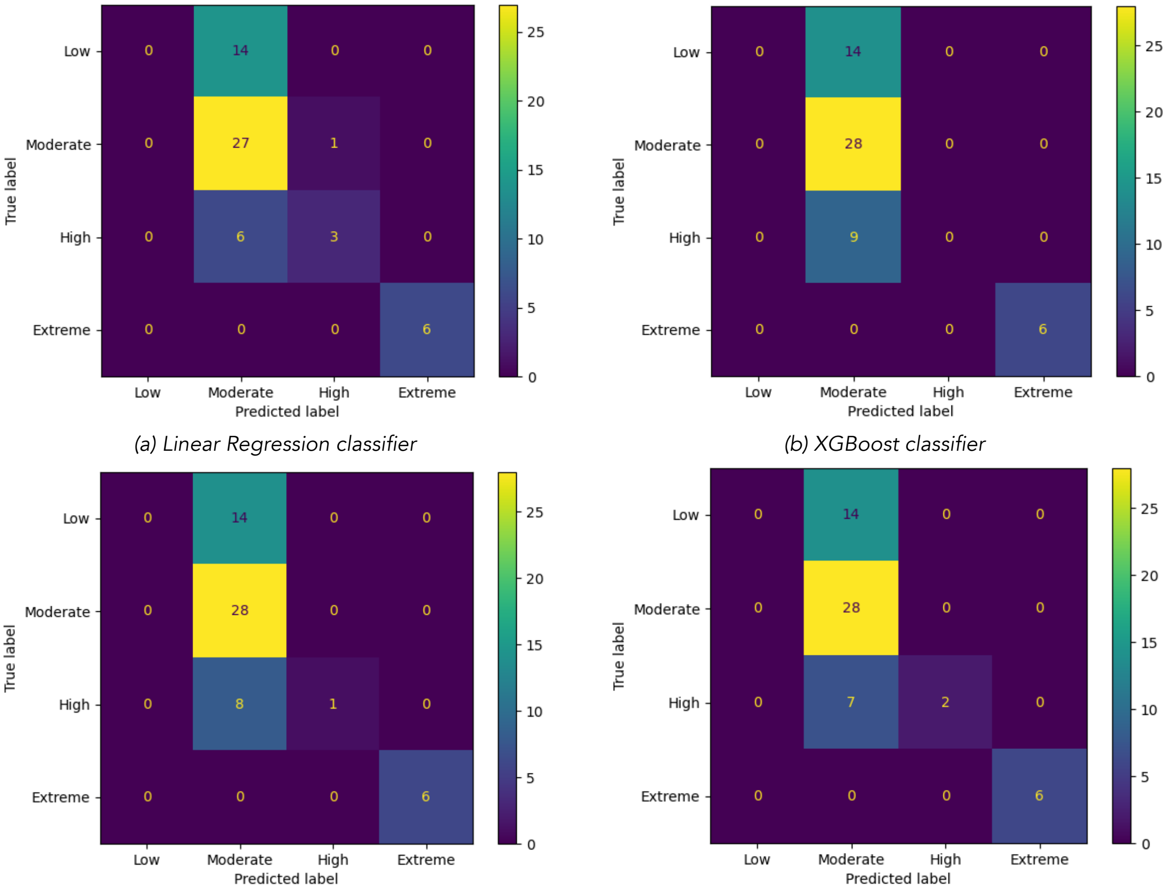


Figure 4-3 – SVM Confusion Matrix over NOA's physical dataset



(a) Linear Regression classifier

(b) XGBoost classifier

(c) Decision Tree classifier

(d) Random Forest classifier

Figure 4-4 – Confusion Matrices for classifiers (a) LR, (b) XGBoost, (c) DT and (d) RF over NOA physical attack scenarios

4.2.1.2. Cyber-attack scenarios

The 46 out of 230 annotated cyber scenarios have been utilised in order to evaluate the performance of the ML models. The results are illustrated in the following table (Table 4-2). The SVM with linear kernel has exhibited the best performance in terms of the accuracy and F1-Score upon the testing dataset. The Random Forest classifier had achieved the 2nd performance overcoming the linear regression classifier. The other 3 classifiers, namely Linear Regression, Decision Tree and XGBoost have quite equal performance. In the following figures, Figure 4-5 and Figure 4-6, the confusion matrix for the SVM and the other classifiers are presented. The classifiers managed to classify correctly the low severity events as well as the majority of the instances for each one of the severity levels.

Classifier	Best set of parameters	Precision	Recall	F1 Score	Accuracy
LR (Ridge)	(alpha=1, normalize=True, tol=0.001, solver="auto", random_state=42)	86.64%	84.78%	84.23%	84.78%
SVM	(kernel="linear", C=0.1, gamma=1, random_state=42)	90.02%	89.13%	**88.54%**	**89.13%**

RF	(criterion="gini", max_depth=110, max_features="auto", min_samples_leaf=10, min_samples_split=10, n_estimators=100, random_state=42, bootstrap=True)	88.27%	86.95%	86.42%	86.95%
DT	(criterion="gini", max_depth=1, max_leaf_nodes=2, min_samples_split=2, random_state=42)	86.64%	84.78%	84.22%	84.78%
XGBoost	(colsample_bytree=0.7, learning_rate=0.01, max_depth=2, min_child_weight=10, n_estimators=80, nthread=4, subsample=0.8)	86.64%	84.78%	84.22%	84.78%

Table 4-2– Performance of the ML Classifiers over NOA's cyber dataset

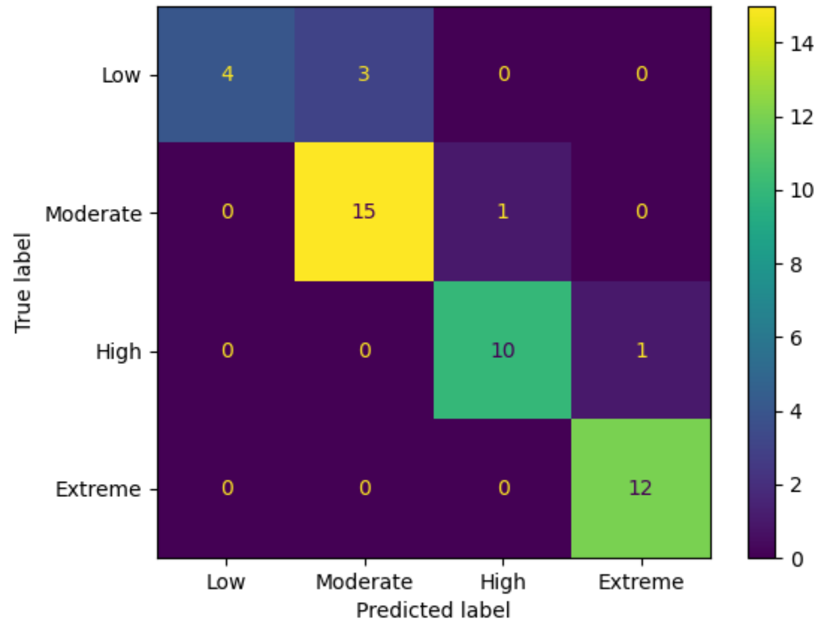
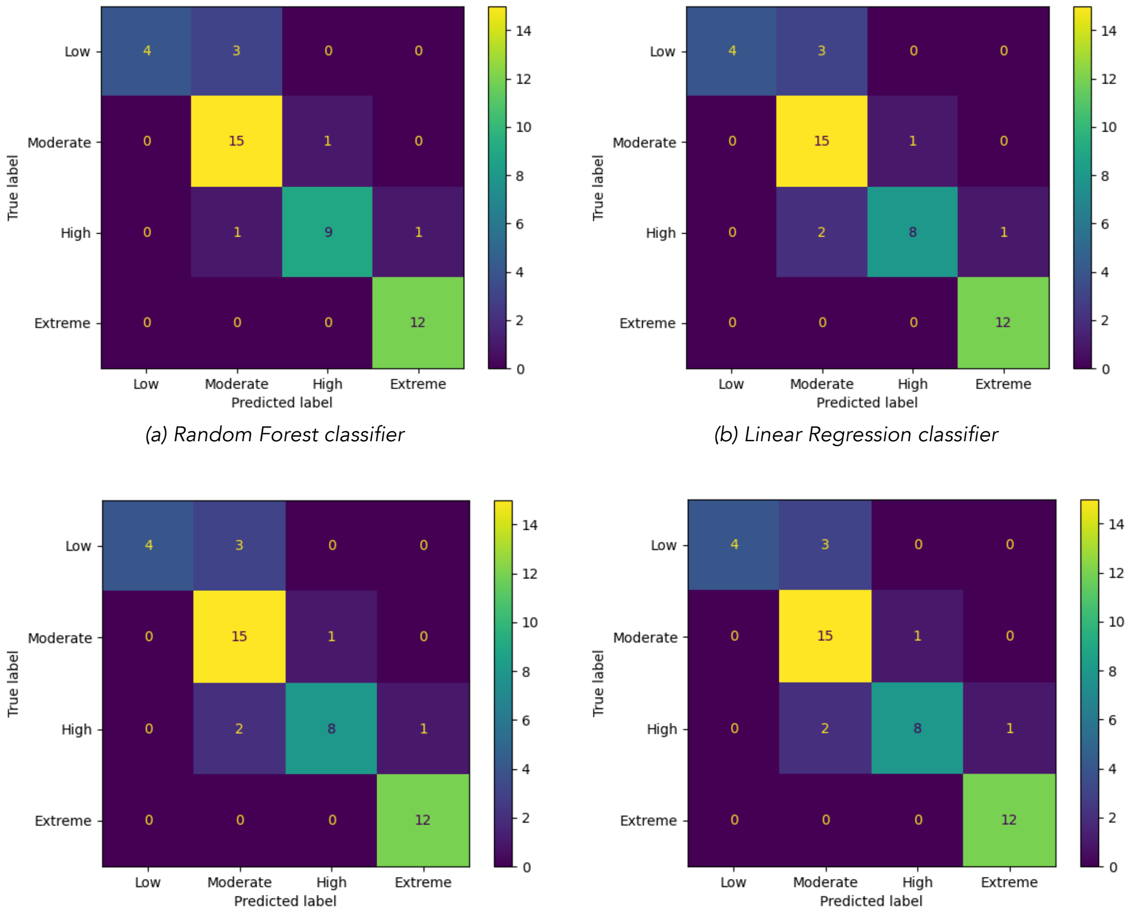


Figure 4-5 – SVM Confusion Matrix over NOA's cyber dataset



(a) Random Forest classifier

(b) Linear Regression classifier

(c) Decision Tree classifier

(d) XGBoost classifier

Figure 4-6 – Confusion Matrices for classifiers (a) RF, (b) LR, (c) DT and (d) XGBoost over NOA cyber-attack scenarios

4.2.2. Experimental Results Over the DEIMOS Pilot

4.2.2.1. Physical attack scenarios

The annotated dataset consists of 301 physical attack hypothetical scenarios divided into 80% (240) events that utilised for the classifiers training and the rest portion of them, 20% (61) events for testing and evaluation purposes. Decision Tree classifier along with Linear Regression exhibit the two highest accuracy rates reaching 60% and 55% correspondingly. The difficulties that appear to classifiers to correctly classify the attack events in terms of their severity level are depicted in the confusion matrices of the classifiers

Classifier	Best set of parameters	Precision	Recall	F1 Score	Accuracy
LR (Ridge)	(alpha=0.1, normalize=True, tol=0.001, solver="auto", random_state=42)	51.19%	58.33%	49.97%	58.33%
SVM	(kernel="linear", C=1, gamma=1, random_state=42)	44.34%	48.33%	45.98%	48.33%

RF	(criterion="gini", max_depth=40, max_features="auto", min_samples_leaf=2, min_samples_split=2, n_estimators=100, random_state=42, bootstrap=True)	51.20%	55.00%	52.64%	55.00%
DT	(criterion="gini", max_depth=1, max_leaf_nodes=2, min_samples_split=2, random_state=42)	44.22%	60.00%	50.71%	60.00%
XGBoost	(colsample_bytree=0.7, learning_rate=0.05, max_depth=2, min_child_weight=1, n_estimators=50, nthread=4, subsample=0.8)	39.26%	50.00%	43.97%	50.00%

Table 4-3– Performance of the ML Classifiers over DEIMOS's physical dataset

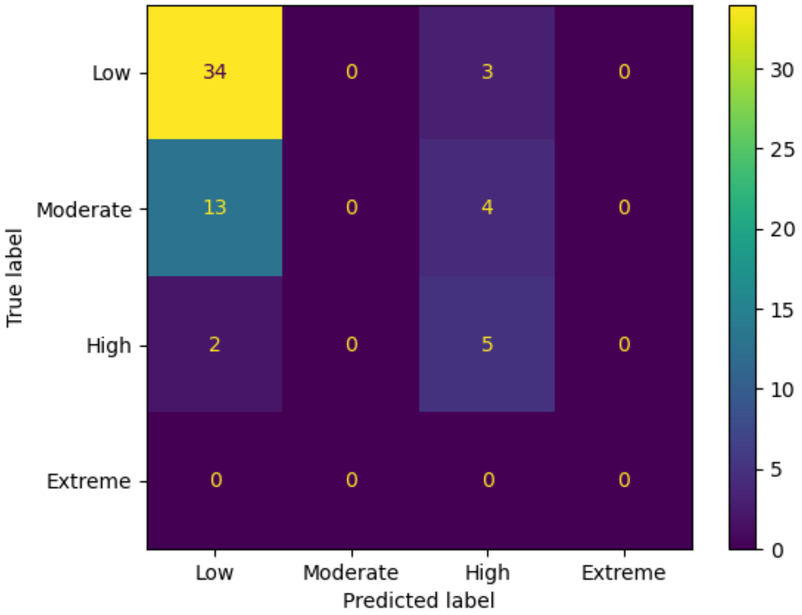
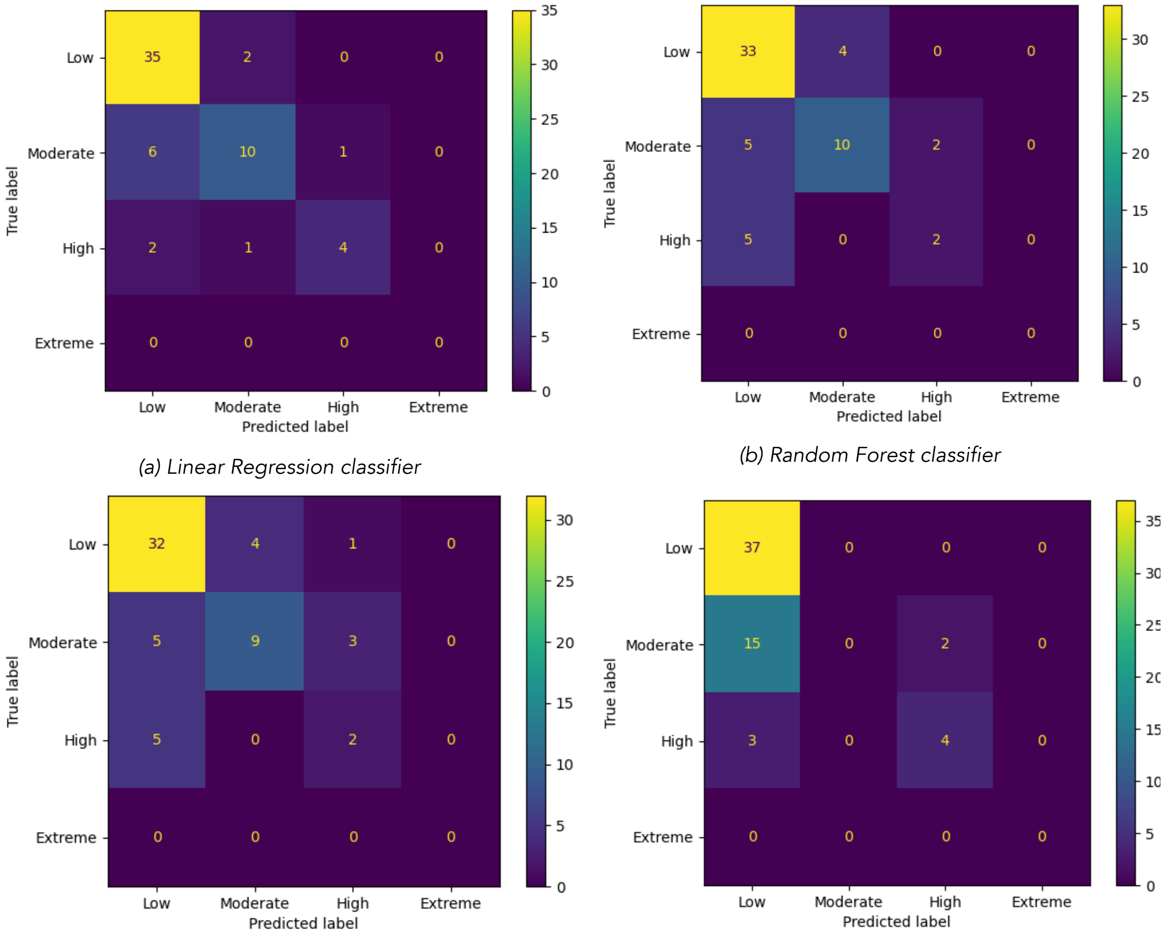


Figure 4-7 – Decision Tree Confusion Matrix over DEIMOS's physical dataset



(a) Linear Regression classifier

(b) Random Forest classifier

(c) XGBoost classifier

(d) SVM classifier

Figure 4-8 – Confusion Matrices for classifiers (a) LR, (b) RF, (c) XGBoost and (d) SVM over DEIMOS physical attack scenarios

4.2.2.2. Cyber attack scenarios

The 458 characterised physical attack scenarios have been divided into a training set that contains 366 entries and a testing set which includes 92 scenarios. In the following table (Table 4-4) the performance of each classifier has been illustrated. Although the Decision Tree is the best performance classifier in terms of the achieved accuracy (~70.65%), however, the other classifiers exhibit similar performance too (~69.56%).

Classifier	Best set of parameters	Precision	Recall	F1 Score	Accuracy
LR (Ridge)	(alpha=0.1, normalize=True, tol=0.001, solver="auto", random_state=42)	69.41%	69.56%	66.63%	69.56%
SVM	(kernel="linear", C=1, gamma=1, random_state=42)	72.52%	69.56%	67.00%	69.56%

RF	(criterion="gini", max_depth=40, max_features="auto", min_samples_leaf=2, min_samples_split=2, n_estimators=100, random_state=42, bootstrap=True)	72.52%	69.56%	67.00%	69.56%
DT	(criterion="gini", max_depth=1, max_leaf_nodes=2, min_samples_split=2, random_state=42)	71.17%	70.65%	67.59%	70.65%
XGBoost	(colsample_bytree=0.7, learning_rate=0.05, max_depth=2, min_child_weight=1, n_estimators=50, nthread=4, subsample=0.8)	69.40%	69.56%	66.63%	69.56%

Table 4-4– Performance of the ML Classifiers over DEIMOS's cyber dataset

From the confusion matrices (Figure 4-9 and Figure 4-10), we can conclude that the DT classifier attained classify correctly the majority of the events with the High severity level, while it has not achieved to classify the scenarios with a moderate severity level.

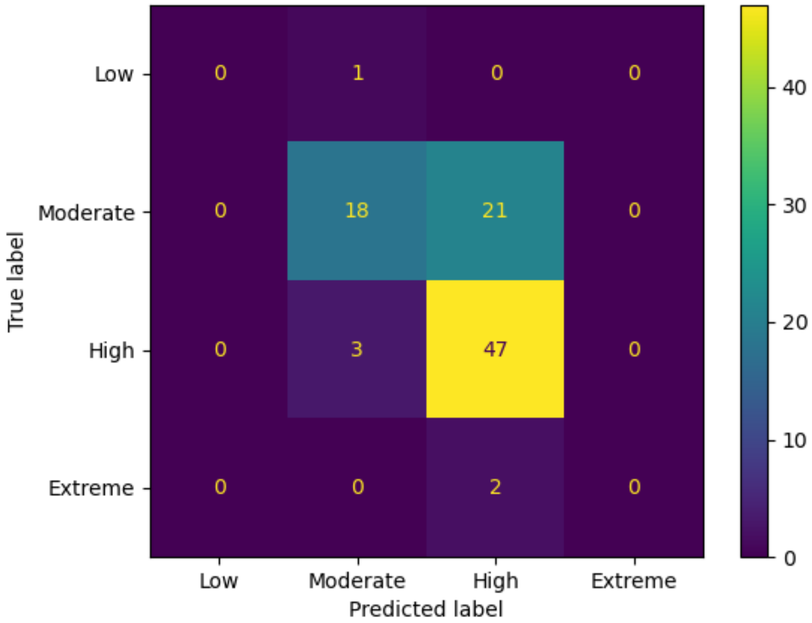
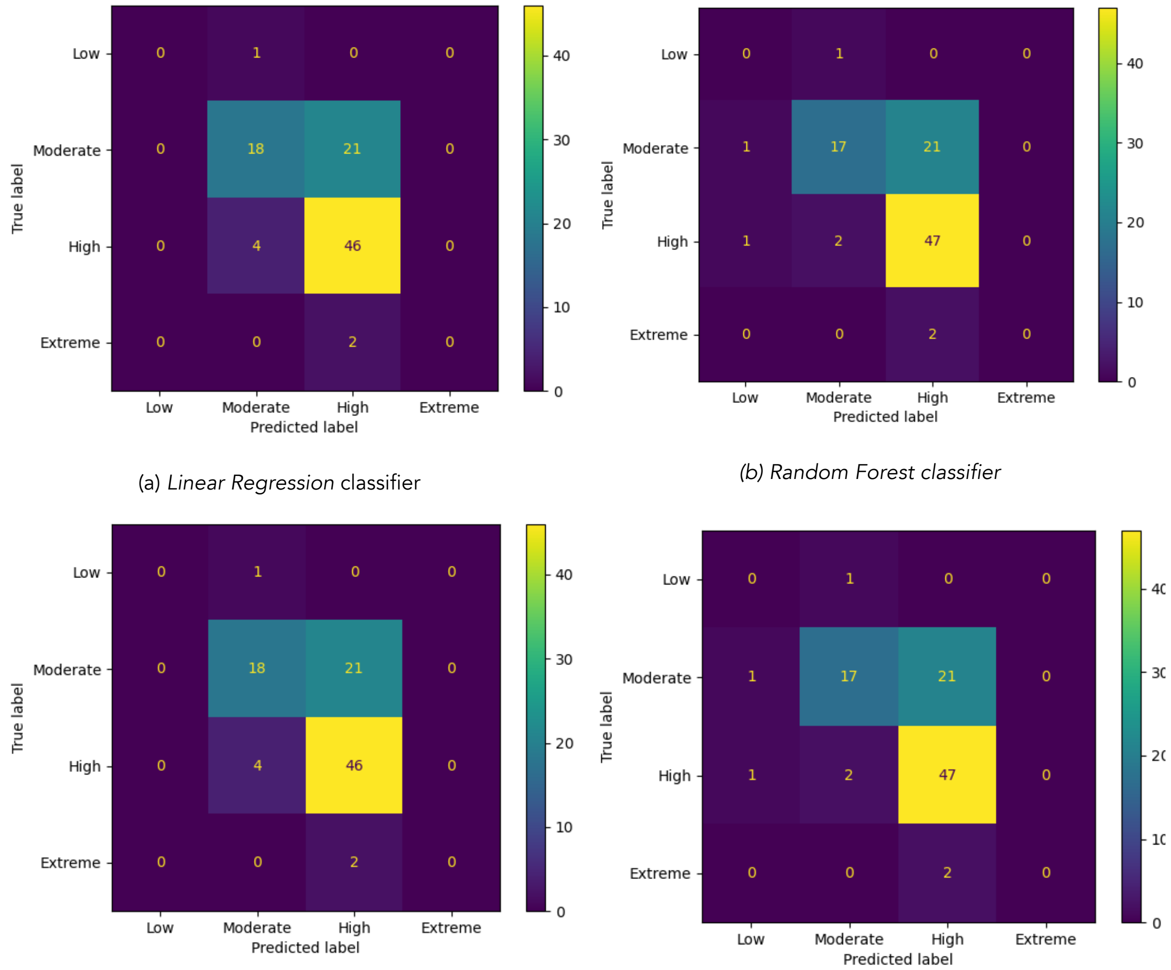


Figure 4-9- Decision Tree Confusion Matrix over DEIMOS's cyber dataset



(a) *Linear Regression* classifier

(b) Random Forest classifier

(c) XGBoost classifier

(d) SVM classifier

Figure 4-10 – Confusion Matrices for classifiers (a) LR, (b) RF, (c) XGBoost and (d) SVM over DEIMOS cyber-attack scenarios

4.2.3. Experimental Results Over the FMI Pilot

In this series of experiments, only physical attack scenarios were annotated and analysed by ML algorithms. The dataset of 202 annotated hypothetical scenarios has been divided into 162 entries for training the models while the rest 40 entries to evaluate (test) the performance of the classifiers. Linear Regression and Random Forest classifiers exhibited the best performance around 61% and 58.5% correspondingly (Table 4-5).

Classifier	Best set of parameters	Precision	Recall	F1 Score	Accuracy
LR (Ridge)	(alpha=1, normalize=True, tol=0.001, solver="auto", random_state=42)	54.08%	60.98%	57.28%	60.98%
SVM	(kernel="linear", C=1, gamma=1, random_state=42)	45.85%	43.90%	44.79%	43.90%

RF	(criterion="gini", max_depth=10, max_features="auto", min_samples_leaf=2, min_samples_split=10, n_estimators=500, random_state=42, bootstrap=True)	51.43%	58.54%	54.75%	58.53%
DT	(criterion="entropy", max_depth=20, max_leaf_nodes=34, min_samples_split=3, random_state=42)	48.64%	36.59%	39.69%	36.59%
XGBoost	(colsample_bytree=0.7, learning_rate=0.01, max_depth=2, min_child_weight=1, n_estimators=50, nthread=4, subsample=0.8)	37.94%	46.34%	40.93%	46.34%

Table 4-5– Performance of the ML Classifiers over FMI's physical attack dataset

In Figure 4-11 the LR confusion matrix is presented. The majority of the attack scenarios that have moderate and high severity levels have been classified correctly. The RF, XGBoost and SVM have been achieved to classify the majority of the scenarios with moderate severity level correctly. However, the classifiers' behaviour has been altered and only the SVM classifier managed to classify correctly the scenarios with high severity level (Figure 4-12). DT classifier managed to classify correctly scenarios of high severity levels while it does not succeed in moderate scenarios.

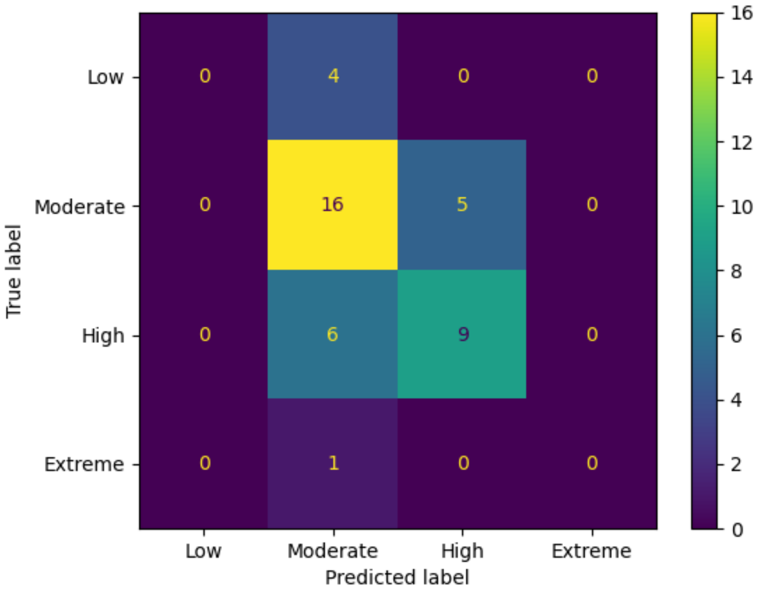
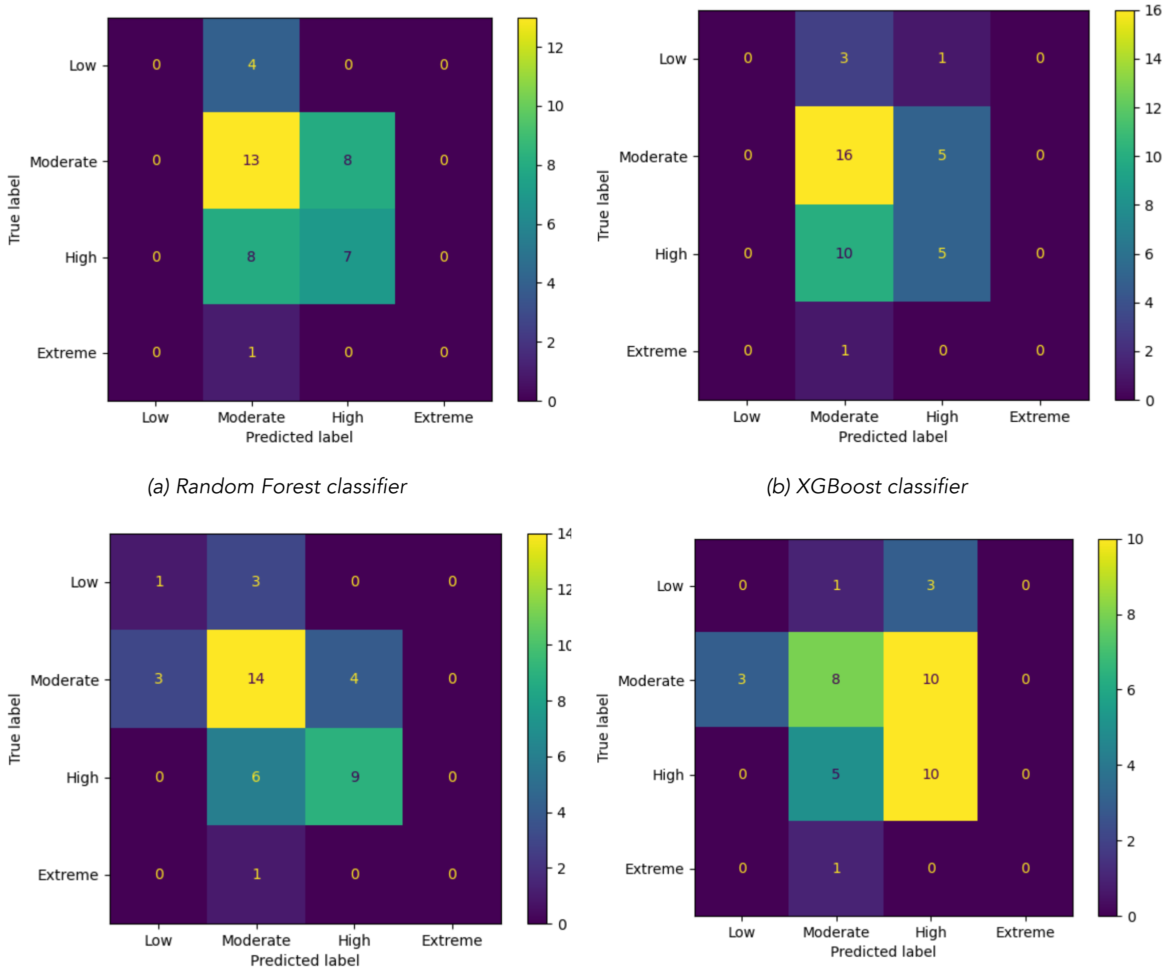


Figure 4-11- LR Confusion Matrix over FMI's physical dataset



(a) Random Forest classifier

(b) XGBoost classifier

(c) SVM classifier

(d) Decision Tree classifier

Figure 4-12 – Confusion Matrices for classifiers (a) Random Forest, (b) XGBoost, (c) SVM and (d) DT over FMI physical attack scenarios

4.2.4. Experimental Results Over the SPACEAPPS Pilot

In this series of experiments, only cyber-attack scenarios were annotated and analysed by ML algorithms. The dataset of 200 annotated hypothetical scenarios has been divided into 160 entries for training the models while the rest 40 entries to evaluate (test) the performance of the classifiers. Linear Regression, Random Forest and Decision Tree classifiers exhibited the best performance around 95% (Table 4-6). Also, the performance of the SVM is high above 87%.

Classifier	Best set of parameters	Precision	Recall	F1 Score	Accuracy
LR (Ridge)	(alpha=1, normalize=True, tol=0.001, solver="auto", random_state=42)	95.43%	95.00%	94.97%	95.00%
SVM	(kernel="linear", C=1, gamma=1, random_state=42)	94.75%	87.50%	90.95%	87.50%

RF	(criterion=”gini”, max_depth=10, max_features=”auto”, min_samples_leaf=2, min_samples_split=10, n_estimators=500, random_state=42, bootstrap=True)	95.43%	95.00%	94.97%	95.00%
DT	(criterion=”entropy”, max_depth=20, max_leaf_nodes=34, min_samples_split=3, random_state=42)	95.43%	95.00%	94.97%	95.00%
XGBoost	(colsample_bytree=0.7, learning_rate=0.01, max_depth=2, min_child_weight=1, n_estimators=50, nthread=4, subsample=0.8)	81.95%	72.50%	69.75%	72.50%

Table 4-6– Performance of the ML Classifiers over SpaceAPPS’s cyber-attack dataset

The classifiers, apart from the XGBoost, had identified correctly the attack scenarios which have moderate severity level. Similar all the classifiers have managed to classify scenarios with high severity level (Figure 4-13 and Figure 4-14).

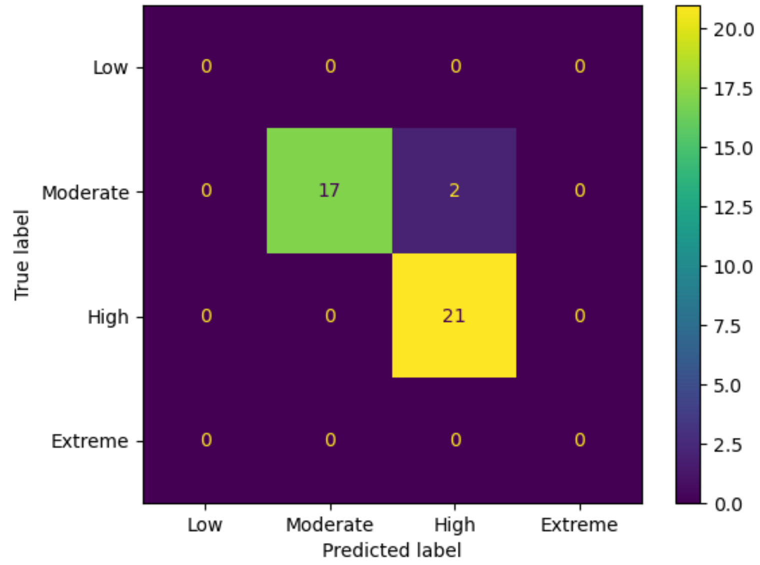
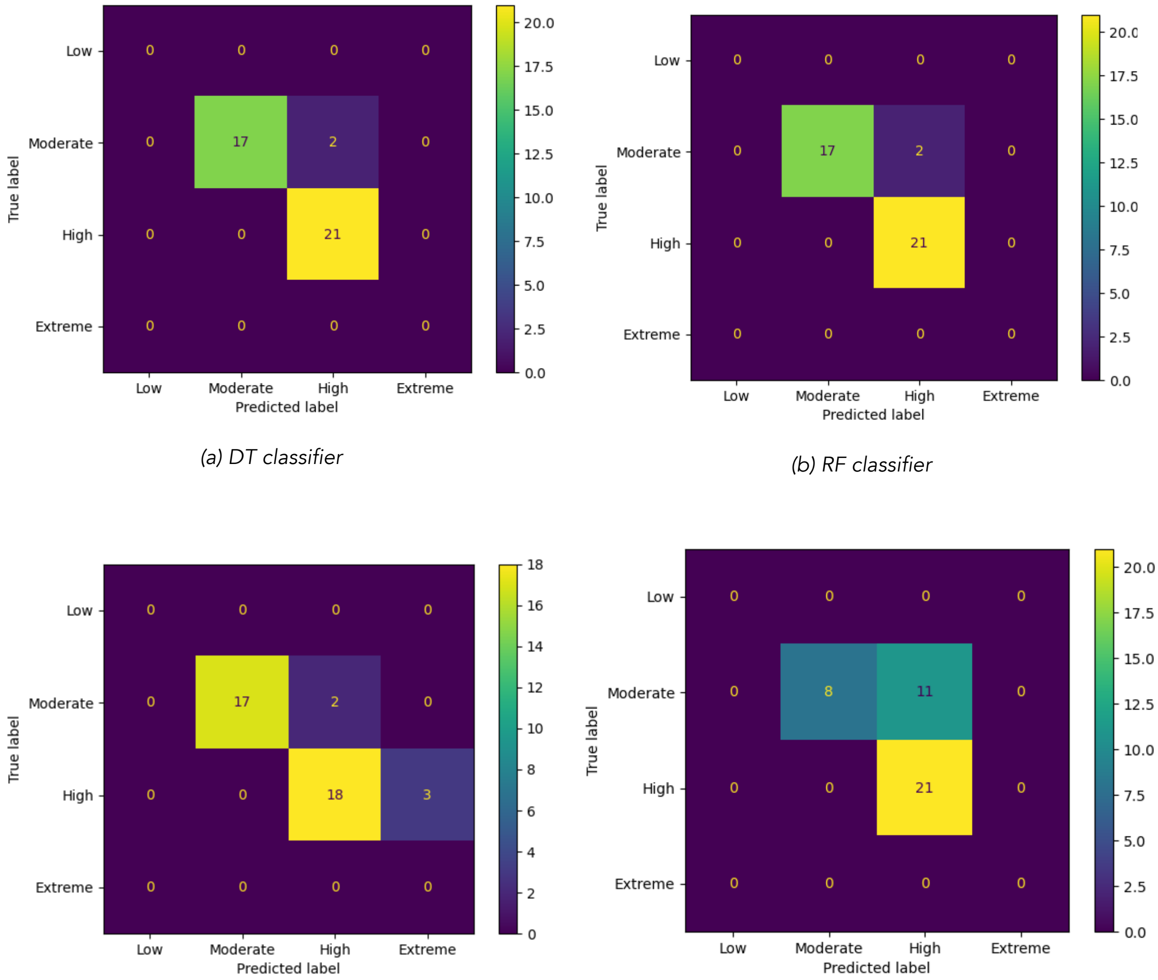


Figure 4-13- Linear Regression Confusion Matrix over SpaceAPPs’ cyber dataset



(a) DT classifier

(b) RF classifier

(c) SVM classifier

(d) XGBoost classifier

Figure 4-14 – Confusion Matrices for classifiers (a) DT, (b) RF, (c) SVM and (d) XGBoost over SpaceAPPS cyber-attack scenarios

4.2.5. Experimental Results Over the SERCO Pilot

In this series of experiments, only cyber-attack scenarios were annotated and analysed by ML algorithms. The dataset of 200 annotated hypothetical scenarios has been divided into 160 entries for training the models while the rest 40 entries to evaluate (test) the performance of the classifiers. Linear Regression, Random Forest and Decision Tree classifiers exhibited the best performance around 95% (Table 4-7). Also, the performance of the SVM is high, slightly above 87%. XGBoost follows with a 72.5% accuracy.

Classifier	Best set of parameters	Precision	Recall	F1 Score	Accuracy
LR (Ridge)	(alpha=1, normalize=True, tol=0.001, solver=”auto”, random_state=42)	92.62%	95.00%	93.78%	95.00%
SVM	(C=0.1, gamma=1, kernel='linear', random_state=42)	92.62%	95.00%	93.78%	95.00%

RF	(criterion="gini", max_depth=10, max_features="auto", min_samples_leaf=2, min_samples_split=10, n_estimators=500, random_state=42, bootstrap=True)	92.62%	95.00%	93.78%	95.00%
DT	(criterion="entropy", max_depth=20, max_leaf_nodes=34, min_samples_split=3, random_state=42)	92.62%	95.00%	72.57%	95.00%
XGBoost	(colsample_bytree=0.7, learning_rate=0.01, max_depth=2, min_child_weight=1, n_estimators=50, nthread=4, subsample=0.8)	62.93%	72.50%	65.60%	72.50%

Table 4-7– Performance of the ML Classifiers over Serco's cyber-attack dataset

Linear Regression, SVM, Decision Tree and Random Forest classifiers exhibit similar behaviour in terms of their performance. They can correctly identify most of the low, moderate and high severity attack scenarios. Last, XGBoost, while predicting most of the moderate and high severity attacks correctly, failed to identify a number of low and high severity scenarios (Figure 4-15 and Figure 4-16).

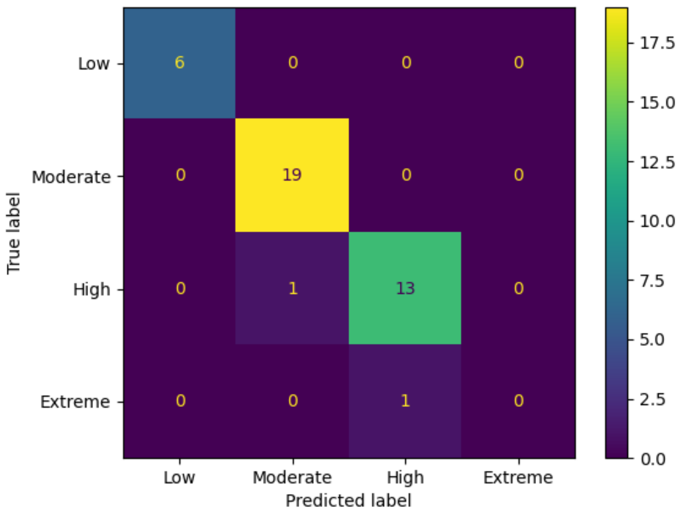
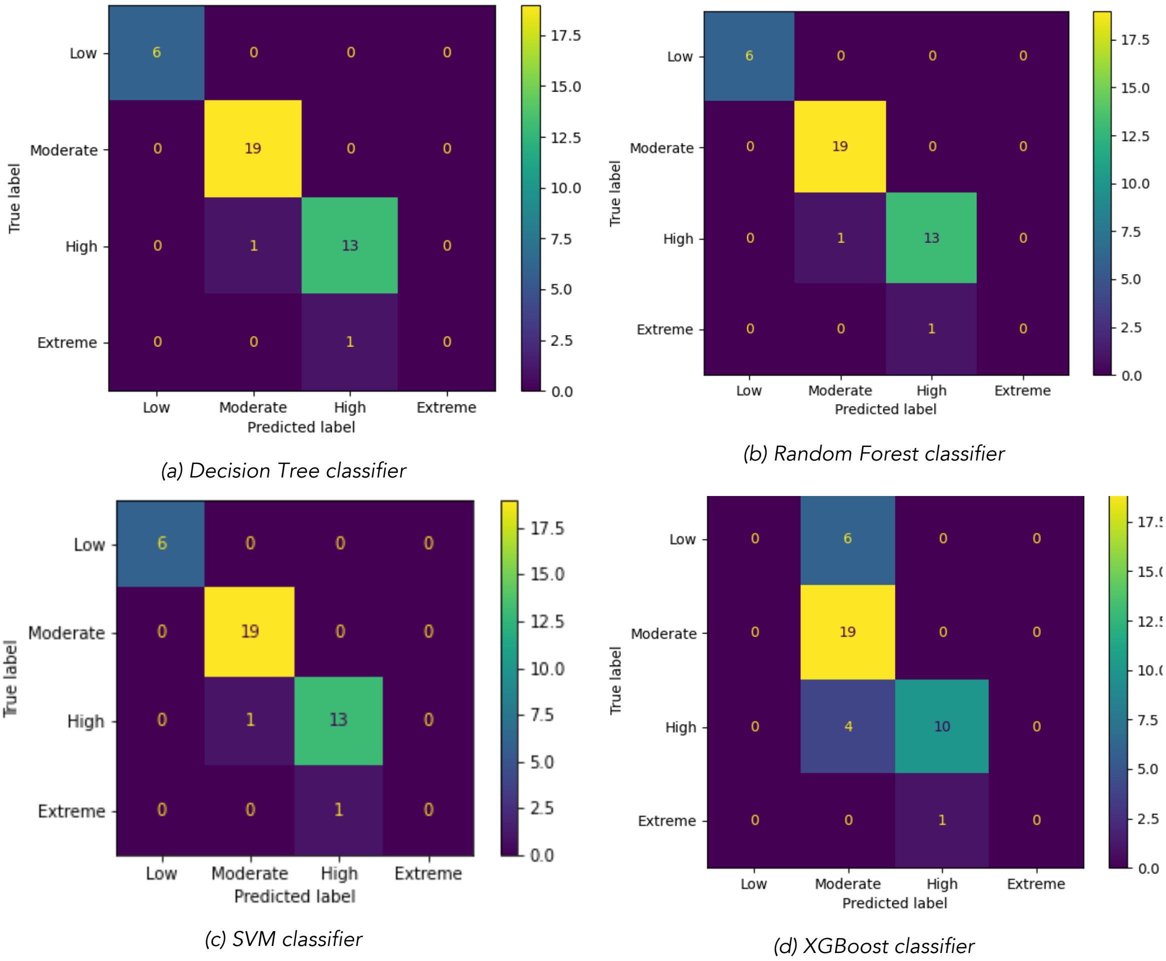


Figure 4-15- Linear Regression Confusion Matrix over Serco's cyber dataset



(a) Decision Tree classifier

(b) Random Forest classifier

(c) SVM classifier

(d) XGBoost classifier

Figure 4-16 – Confusion Matrices for classifiers (a) DT, (b) RF, (c) SVM and (d) XGBoost over Serco cyber-attack scenarios

5. Conclusions and Future Outlook

In the present deliverable, we detailed the Security Risk Assessment framework that enables the dynamical estimation of the severity level of physical, cyber and complex malicious events that take place in Ground Segments of Satellite assets. The framework consists of two major components, namely the Information Fusion module and the Decision Fusion module. The former utilises machine learning approaches to analyse the correlated physical, cyber and hybrid events. The latter receives the assessments of the severity level from the previous analysis and updates the situational picture dynamically. For the training of the machine learning models require the utilisation of annotated datasets that contain classified P/C events in terms of the level of severity. For this purpose, the involvement of the experts and operators of the domain of the Satellite Ground Segments was requested. To facilitate this process a dedicated web-based tool, called Annotation Tool, was created and deployed to end-users. Then, machine learning methods were applied to the annotated datasets and evaluated against well-known validation measures. The best ML models were chosen, so as to create a suite of tools per pilot site and type of event (physical or cyber), which is enabled to classify "unknown" events. Furthermore, in the next level (decision fusion) the severity level estimations can be further enhanced based on specific rulesets formed by experts and combined so as to provide the overall severity assessment for complex threatening events.

In general, from the analysis of the experimental results, we can conclude that the proposed framework is quite robust. During the operational test on NOA premises, its application can be judged as reliable providing real-time assessments covering the operators' specifications. The proposed framework will further evaluate during the operational and demo tests in the various pilot sites at the 7SHIELD project.

However, some limitations have been identified concerning its performance. Specifically, the training of the machine learning models depends on the adequate annotated datasets. Hence, the diversity of the datasets that cover various complex instances is a crucial driver to improve the performance of the training. Also, sophisticated ensemble machine learning methods can be applied to exploit the benefits of independent classifiers.

6. References

[1] Setola, R., Luiijf, E., Theocharidou, M. (2016). Critical Infrastructures, Protection and Resilience. In: Setola, R., Rosato, V., Kyriakides, E., Rome, E. (eds) Managing the Complexity of Critical Infrastructures. Studies in Systems, Decision and Control, vol 90. Springer, Cham. https://doi.org/10.1007/978-3-319-51043-9_1

[2] European Council (2008) Council Directive 2008/114/EC of 8 December 2008 on the identification and designation of European critical infrstructures and the assessment of the need to improve their protection (Text with EE Arelevance), Brussels, Dec 2008. Available online at http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32008L0114&from=EN. Retrieved on 27 Oct 2016

[3] EC: Overview of natural and man-made disaster risks the European union may face (2020). https://ec.europa.eu/echo/sites/default/files/overview_of_natural_and_man-made_disaster_risks_the_european_union_may_face.pdf

[4] EY: Evaluation study of council directive 2008/114 on the identification and designation of european critical infrastructures and the assessment of the need to improve their protection (2019). https://op.europa.eu/en/publication-detail/-/publication/118dcd3d-b041-11ea-bb7a-01aa75ed71a1. https://doi.org/10.2837864404

[5] Baubion, C. (2013). OECD Risk Management: Strategic crisis management. DOI: https://doi.org/10.1787/5k41rbd1lzr7-en

[6] United Nations International Strategy for Disaster Reduction (2019). Global Assessment Report on Disaster Risk Reduction 2019. United Nations. https://doi.org/10.18356/f4ae4888-en, https://www.un-ilibrary.org/content/books/9789210041805

[7] Hegde, J., Rokseth, B. (2020). Applications of machine learning methods for engineering risk assessment - a review. Safety Sci. 122, 104492. https://doi.org/10.1016/j.ssci.2019.09.015

[8] Makri, R.; Karaivazoglou, P.; Kyritsis, A.; Skitsas, M.; Koutras, N.; Valera, J.; Sanchez, J.M. (2020). Modern innovative detectors of physical threats for Critical Infrastructures. In Cyber-Physical Threat Intelligence for Critical Infrastructures Security: A Guide to Integrated Cyber-Physical Protection of Modern Critical Infrastructures; Soldatos, J.; Philpot, J.; Giunta, G., Eds. Boston–Delft: Now Publishers, pp. 397–414. doi:10.1561/9781680836875.ch22

[9] Hutter, D. (2016). The importance of physical security in the workplace. https://resources.infosecinstitute.com/topic/importance-physical-security-workplace/

[10] Choraś, M., Kozik, R., Flizikowski, A., Hołubowicz, W., Renk, R. (2016). Cyber Threats Impacting Critical Infrastructures. In: Setola, R., Rosato, V., Kyriakides, E., Rome, E. (eds) Managing the Complexity of Critical Infrastructures. Studies in Systems, Decision and Control, vol 90. Springer, Cham. https://doi.org/10.1007/978-3-319-51043-9_7

[11] Sarker, I.H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. SN COMPUT. SCI. 2, 160. https://doi.org/10.1007/s42979-021-00592-x

[12] Antzoulatos, G.; Giannakeris, P.; Koulalis, I.; Karakostas, A.; Vrochidis, S.; Kompatsiaris, I. A Multi-Layer Fusion Approach For Real-Time Fire Severity Assessment Based on Multimedia Incidents. Proceedings of the 17th Int. Conf. on Information Systems for Crisis Response and Management (ISCRAM'20), eds. Amanda Lee Hughes, Fiona McNeil and Christopher Zobel, 24-27 May 2020. [*zenodo link*]

[13] Antzoulatos, G.; Karakostas, A.; Vrochidis, S.; Kompatsiaris, I. The Crisis Classification component to strengthen the early warning, risk assessment and decision support in extreme climate events. Ilias S. Kotsireas, Anna Nagurney, Panos M. Pardalos, Arsenios Tsokas (eds). Dynamic of Disaster - Impact, Risk, Resilience, and Solutions, Springer, 2021, pp. 39–66. DOI: https://doi.org/10.1007/978-3-030-64973-9.

[14] Antzoulatos, G.; Koulalis, I.; Karakostas, A.; Vrochidis, S.; Kompatsiaris, I. Towards to integrate a multi-layerMachine Learning Data Fusion approach into Crisis Classification and Risk Assessment of extreme natural events. Akhgar, Babak and Kavallieros, Dimitrios and Sdongos, Evangelos (eds) Technology Development for Security Practitioners. Book series: Security Informatics and Law Enforcement, Springer, 2021, pp. 513–537. DOI: https://doi.org/10.1007/978-3-030-69460-9_30.

[15] Antzoulatos, G.; Kouloglou, I.; Bakratsas, M.; Moumtzidou, A.; Gialampoukidis, I.; Karakostas, A.; Lombardo, F.; Fiorin, R.; Norbiato, D.; Ferri, M.; Symeonidis, A.; Vrochidis, S.; Kompatsiaris, I. Flood Hazard and Risk Mapping by Applying an Explainable Machine Learning Framework Using Satellite Imagery and GIS Data. Special Issue: Environmental Water Monitoring for Sustainable Development in Urban and Rural Areas, Sustainability, 2022, Volume 14, Issue 6. DOI: https://doi.org/10.3390/su14063251.

[16] Antzoulatos, G.; Orfanidis, G.; Giannakeris, P.; Tzanetis, G.; Kampilis-Stathopoulos, G.; Kopalidis, N.; Gialampoukidis, I.; Vrochidis, S. and Kompatsiaris I. Severity Level Assessment from Semantically Fused Video Content Analysis for Physical Threat Detection in Ground Segments of Space Systems. In: Katsikas S. et al. (eds) Computer Security. ESORICS 2021 International Workshops. ESORICS 2021. Lecture Notes in Computer Science, 2022, vol 13106. Springer, Cham. DOI: https://doi.org/10.1007/978-3-030-95484-0_27.

[17] Wagenaar, D., Curran, A., Balbi, M., Bhardwaj, A., Soden, R., Hartato, E., Mestav Sarica, G., Ruangpan, L., Molinario, G., Lallemant, D. (2020). Invited perspectives: How machine learning will change flood risk and impact assessment. Nat. Hazards Earth Syst. Sci., 20(4), pg 1149-1161. Copernicus Publications. DOI: 10.5194/nhess-20-1149-2020

[18] GFDRR (2018). Machine Learning for Disaster Risk Management, GFDRR, Washington, D.C., USA.

[19] Said, N., Ahmad, K., Riegler, M., Pogorelov, K., Hassan, L., Ahmad, N., Conci, N. (2019). Natural disasters detection in social media and satellite imagery: a survey. Multimedia Tools and Applications 78(22), pp. 31267-31302. DOI: 10.1007/s11042-019-07942-1

[20] Yu, M., Yang, C., Li, Y. (2018). Big data in natural disaster management: A review. Geosciences 8(5). DOI: 10.3390/geosciences8050165

Horizon 2020
European Union Funding
for Research & Innovation

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 883284