Preprints.org

Article

Not peer-reviewed version

The Impact of Data Injection on Predictive Algorithm Developed Within Electrical Manufacturing Engineering in the Context of Aerospace Cybersecurity

Jorge Bautista-Hernández * and María Ángeles Martín-Prats

Posted Date: 27 September 2023

doi: 10.20944/preprints202309.1812.v1

Keywords: predictive algorithms cybersecurity; machine learning; advanced persistent threats

Preprints.org is a free multidiscipline platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 September 2023 doi:10.20944/preprints202309.1812.v1

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

The Impact of Data Injection on Predictive Algorithm Developed within Electrical Manufacturing Engineering in the Context of Aerospace Cybersecurity

Jorge Bautista-Hernández [1,2,*] and María Ángeles Martín-Prats [1]

[1] Department of Electronics Engineering, University of Seville, Spain
[2] Department of Electrical Engineering, Airbus Poland, Warsaw, Poland
* Correspondence: jorbauher@alum.us.

Abstract: Cybersecurity plays a relevant role in the new digital age in aerospace industry. Predictive algorithms are necessary to interconnect complex systems within the cyberspace. In this context, where security protocols do not apply, challenges to maintain data privacy and security arise for the organizations. Thus, the need of cybersecurity is required. The four main categories to classify threats are interruption, fabrication, modification and interception. They all share a common thing, soften the three pillars which cybersecurity needs to guarantee. These pillars are confidentiality, availability and integrity of data (CIA). Data injection can contribute to this event by creation of false indicators which can lead to errors creation during the manufacturing engineering process. In this paper, the impact of data injection on existing dataset used on manufacturing process is shown. The design model synchronizes the following mechanisms developed within machine learning techniques which are, the risk matrix indicator to assess the probability of producing an error, the dendrogram to clusters the dataset in groups with similarities, the logistic regression to predict the potential outcomes and the confusion matrix to analyze the performance of the algorithm. The results presented in this study, which was carried out using a real dataset related to the electrical harnesses installed in a C295 military aircraft, estimate that injection of false data indicators increase the probability of errors creation in 24.22 % on the predicted outcomes required for the generation of the manufacturing process. Overall, implementing cybersecurity measures and advanced methodologies to detect and prevent cyberattacks are necessary.

Keywords: predictive algorithms cybersecurity; machine learning; advanced persistent threats

1. Introduction

Latest reports in 2022 from the European Union Agency for Cybersecurity (ENISA) show 586 reported cybersecurity incidents against 77 in 2012. Cyberattacks are increasing not only in frequency but also in complexity affecting organizations worldwide. Safety is an important pillar to protect the overall assets. Risk assessments procedures are likely to define level of impact, the vulnerabilities and the affected assets in order to minimize the risk and reach the highest level of safety [1]. This situation is not possible due to security protocols and privacy do not apply in the new digital age. On the other hand, predictive algorithms which are developed to automatically perform processes and reduce costs in organizations are sensitive to threats such as for example data modifications. Thus, data protection is essential in this context. Cybersecurity is needed to ensure confidentiality, availability and integrity of data from untrusted sources access. The aim of this study is to analyze the impact of data modification in order to observe the time increased in electrical harnesses manufacturing and error rate as hazard outcomes of the aerospace predictive algorithm after the dataset has been compromised. The predictive algorithm performance is also shown before and after the event has

© 2023 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 September 2023 doi:10.20944/preprints202309.1812.v1

occurred. Thus, countermeasures to protect the dataset in the cybersecurity context are considered and applied for this purpose.

Algorithms development aim to perform tasks faster whereas they are also design to enhance safety. Most of them are modelled to replace manual tasks by automation. Technologies such as machine learning developed within the artificial intelligence are key to develop such as algorithms [2]. Indeed, they are increasing in quantity since the new digital age requires more digital sources in order to interconnect systems to overcome the system complexity in the cyber-physical spaces. On the other hand, multiple users not only have access to the digital application where algorithms are executed and outcomes are display but also, they are frequent users in a daily basis. Due to all of this, data can be compromised. Thus, it is necessary to protect data using techniques within the framework of the cybersecurity.

The four main categories to classify threats types are, interruption, fabrication, modification and interception in the cybersecurity context [3]. They all aim to develop malicious content in a system. Complex techniques developed by attackers and malware means can still be successful to manipulate legitimate of the users what is out of this paper scope [4].

In this paper, an analysis after injecting malicious data impacting on the predictive algorithms performance is performed. This algorithm was previously detailed and was developed within the aerospace industry [5]. We focus on the performance of the algorithm after the dataset is compromised. Analysis of the impact in time and error rate is carried out. The hypothesis we have considered to simplify the results is that the attacker has information of the dataset of the victim and has succeeded to access the system. The following research questions are:

- Can the data injection stop manufacturing of the electrical harness?
- Can this event be avoided by application of proper cyber defence techniques?
- Does the quantity of compromised data affect to the efficiency of the algorithm?

The remainder of this paper is as follows. The review of the related work in section 2, section 3 is the research methodology related to the validation model. The main results are presented in section 4. Discussions are in section 5 and finally to summarize in section 6 conclusions and future work.

2. Materials and Methods

The increasing interconnectivity and complexity of the systems requires to use advanced technologies in the cyberspaces. Cyberspaces are highly relying on digital applications which are based in innovative techniques developed within the context of artificial intelligence such as predictive algorithms. Machine learning techniques used to develop predictive algorithms play a relevant role to be implemented within the organization to optimize manufacturing process, mitigate errors and enhance safety. Cyberspaces are more vulnerable and exposed to cyberattacks. Thus, the importance of cybersecurity has raised as a main priority to maintain the integrity and security of digital cyberspaces [4].

This study aims to compromise the existing dataset which is used as inputs for predictive algorithms developed in the electrical manufacturing in aerospace in order to analyze the algorithm performance and the main consequences related to the error rate within the results after data injection has been carried out. In the cybersecurity context, the way to compromise data can be presented as follows:

- Data modification: Malicious can search for specific data within the dataset and modify it to achieve their goals.
- Changing random values: Change data values randomly to cause confusion and make the data less reliable.
- Data deletion: To delete certain information can cause significant problems, especially if the deletion of the data is critical to the business or customers.
- Data reformatting: Change the data format in order to make it more difficult to use.
- Insertion of false data: False data into the dataset to deceive users who query it.

The research was performed at the electrical harness department in aerospace industry using a dataset related to 157 harnesses installed in a military aircraft C295. In defence sector, the simulation

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 September 2023 doi:10.20944/preprints202309.1812.v1

for data injection could represent a significant risk to the digital assets. The rise of Advanced Persistent Threats (APT) which are a very high sophisticated malware, are evolving aiming to avoid security measures. The attackers often send phishing emails until the first target system gets compromised. Once the malware has been deployed, other intrusive tools can enable the propagation to the internal network. Consequently, data extraction can be conducted, allowing the attackers to steal sensitive information. Due to all these significant risks, is essential for the organization to adopt cybersecurity measures [6].

The aim of this paper is to recognize potential breaches and analyse the consequences in order to strengthen the protection of the digital assets.

The detailed model algorithm, which was developed and is described in previous research paper uses as inputs real dataset used for manufacturing of electrical harness on a C295 aircraft [5]. Once the algorithm has analysed the injected data, the main outcomes which are represented in the following Figure 1 are modified conducting to wrong results.

The general structure of the model is also shown in Figure 1. The Risk matrix calculates the likelihood of error creation during manufacturing process. The automation tool ensures process are automatically generated. The dendrogram is used to group the dataset defining the clusters. The confusion matrix is acting as a controller of the model by predicting inaccurate outputs. The data generation is analysed and processed through the following algorithms:

- Algorithm 1 which shows as the main output the risk matrix.
- Algorithm 2 which uses techniques based on clustering hierarchy agglomerative.
- Algorithm 3 which presents the performance of the model by the confusion matrix.

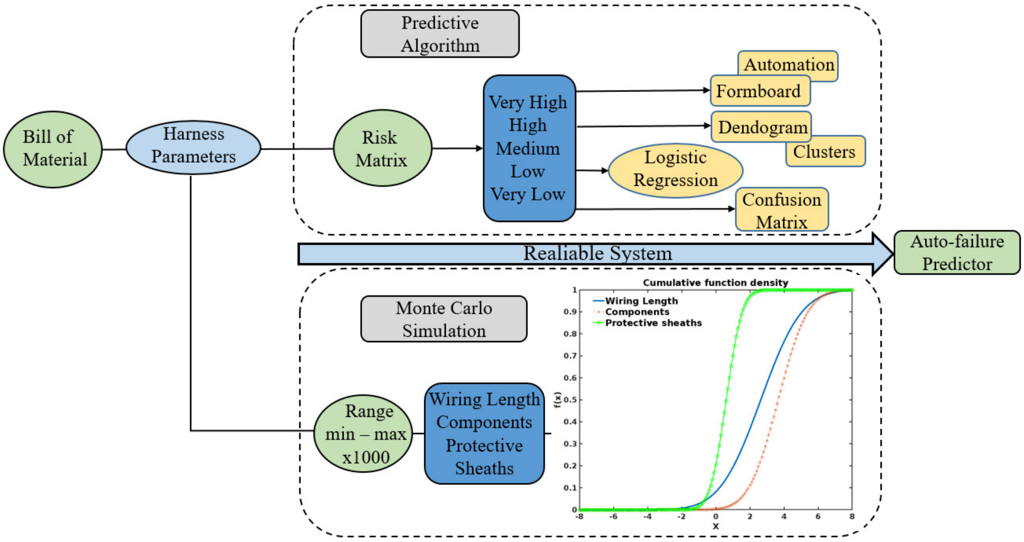

Figure 1. Diagram representation of the structure of the predictive algorithm.

The model was also validated using Monte Carlo techniques which are also described in detail in previous paper [7]. The cumulative function density is presented as the probability function which defines the threshold between the main parameters selected and the likelihood of occurrence. The 1000 runs give enough evidences to consider the predictive algorithm as a reliable system to be implemented as auto failure-predictor.

3. Results

In particular the results are shown in three different blocks in order to assess the impact in the manufacturing process considering the main outcomes indicators.

- Risk Matrix: The assessment performed prior to manufacturing provides the level of risk in error creation during the manufacturing process.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 September 2023 doi:10.20944/preprints202309.1812.v1

- Dendrogram: The grouping of the outcomes in families sharing similarities help to identify patterns and provide good predictions using a logistic regression for new dataset.
- Confusion Matrix: The classification of the results in true positives, true negatives, false positives and false negatives not only help to define the performance of algorithm but also to detect inconsistences within the dataset. This situation is shown by generation of the big amount of true or false negatives.

The Table 1 shows the likelihood of error creation before and after data injection. The real dataset of 157 electrical harness part numbers manufactured for a C295 military transport aircraft show that only 3.18 % of the harness present a high risk in terms of error creation, 18.47 % a moderate risk and most of the harness 78.34 % a low risk of error creation during manufacturing processes. However, after the dataset has been modified, 93.63 % of the harness show a low risk, 6.36 % a moderate risk and none of the harness present a high risk for an error to occur during the process. The probability of error creation during the manufacturing process has increased in 24.22 %. This situation threatens the safety in aerospace. Thus, countermeasures are necessary to be applied and protect the dataset.

Table 1. Likelihood of error creation associate to the risk matrix before and after data injection.

Risk Matrix	Real Data	Injected Data
High	3.18	0
Moderate	18.47	6.36
Low	78.34	93.63

From the Figure 2, it is observed that the impact of data injection has modified the real dataset. This situation has created a false behaviour on the algorithm performance. The risk matrix has established different scores on each part number of electrical harness defining the probability of error creation. The blue line represents the real dataset associated to the real risk matrix for the experimental dataset used and the orange line represents the variation on the risk matrix after data has been injected. This false behaviour generated by the data injected is affecting to the calculation of manufacturing time.

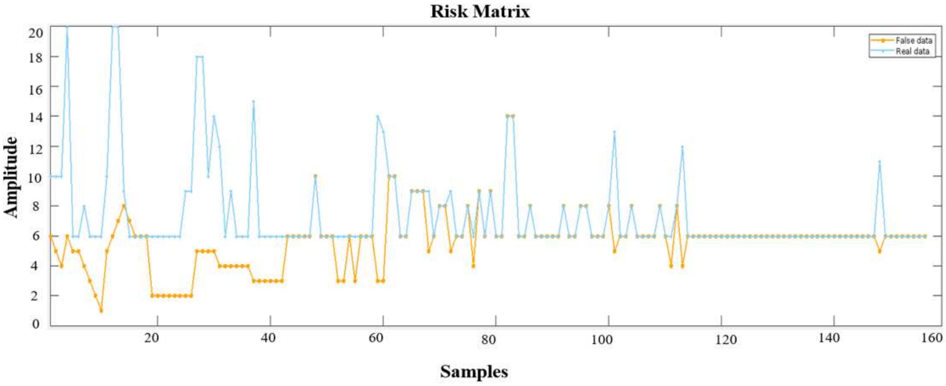

Figure 2. Risk matrix representation of dataset before injection (real data) and after injection (false data).

The Figure 3 shows the total manufacturing time established for the entire real dataset has decreased from 239.3 h in the green bar to 182.5 h represented in the grey bar. The time has decreased in 23.73 % after data injection generating false time calculation. This anomaly can be an indicator of cyberattack.

doi:10.20944/preprints202309.1812.v1

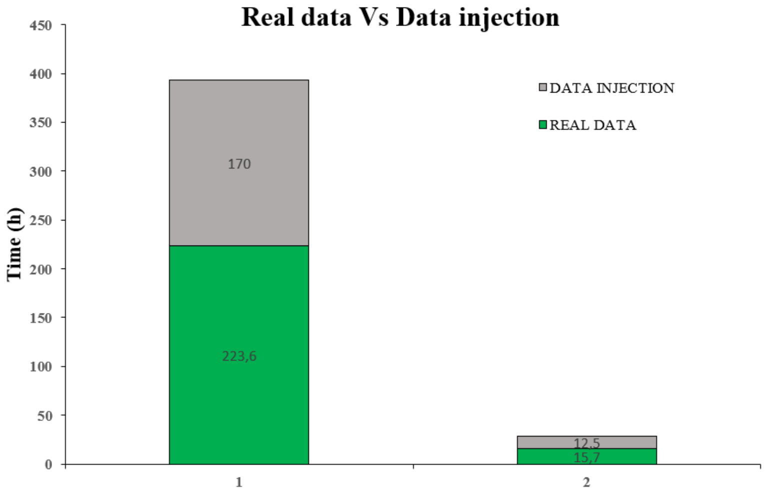

Figure 3. Manufacturing time representation of dataset before injection (real data) and after injection (false data).

Figure 4 represents the hierarchy groups showing similarities between the dataset and grouping them in families. This situation allows to identify patterns and improve the making decisions. The number of clusters have been increased after data injection from 1 cluster to 3. In the case of real dataset in this aircraft is possible to distinguish one family which needs to receive special attention, groups with similarities established are the clusters I, J, K and L, M, N, O, P. The rest of the dataset is compact within the main cluster. However, after data injection, the number of clusters has increased to 3 and the data has been dispersed as shown in Figure 4. After dispersion of the dataset, the following groups are defined from A′ to R′.

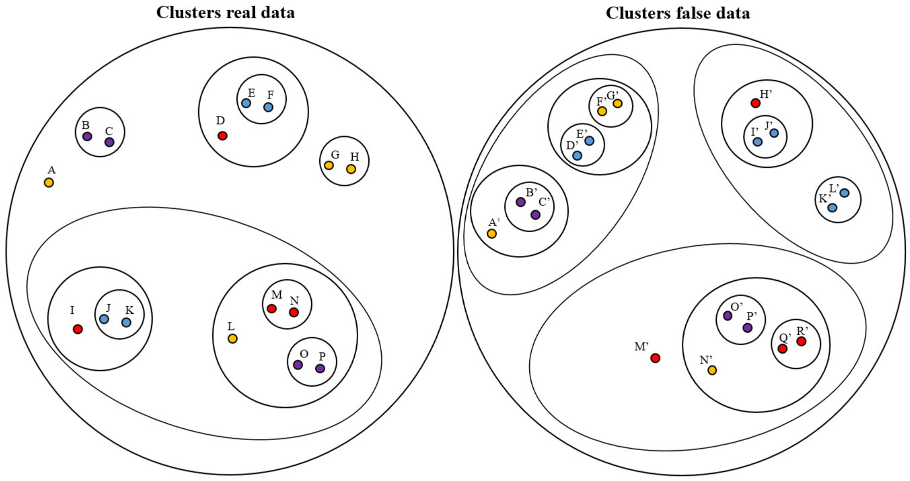

Figure 4. Dendrograms show groups of similar data in each node before injection (real data) and after injection (false data).

Figure 5 represents the confusion matrix showing the performance of the algorithm. After running the simulation with the real dataset, the results are the following: TN = 26, TP =6, FN = FP = 0. However, after data is injected, the results are: TN′ = 1, TP′ =3, FN′ = 0, FP′ = 27. It is observed the increase in the number of false positives. This situation indicates a decrease in the performance of the predictive algorithm. Therefore, the data injected into the dataset can be used as a good indicator for cyberattack detection. The confusion matrix collects all the threats indicators. This early detection mechanism enables fast reaction and reduce the malicious risk [8].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 September 2023 doi:10.20944/preprints202309.1812.v1

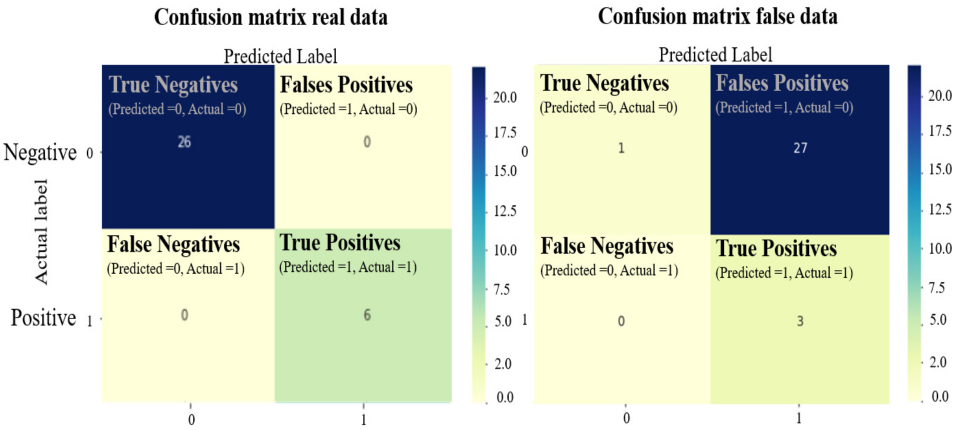

Figure 5. Confusion Matrix outcomes model before injection (real data) and after injection (false data).

The following Table 2 shows the main outcomes metrics defined for the algorithm showing the impact generated in the performance after data has been injected losing model reliability.

Table 2. Metrics calculation to evaluate algorithm performance before and after data injection.

Metrics Comparison	Data Real	Data Injection
$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}}$	1.0	0.1
$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}}$	1.0	1.0
$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}}$	1.0	0.19
$F_1 = \frac{2\ \text{TP}}{2\ \text{TP} + \text{FP} + \text{FN}}$	1.0	0.13

Overall, the data injection compromises the integrity principle of the CIA [9]. As the attacker modifies the inputs, the detection mechanism is necessary to identify the situation. This anomaly behaviour will affect not only to manufacturing time but also to the risk matrix function Z, which shows the probability of creating an error during the manufacturing process and is critical for the correct behaviour of the algorithm. Thus, security practises are essential for ensuring safety and protection of the design model. The protective mechanism to meet this requirement is based on the following detection strategy, if equation (1) is satisfied then the algorithm will continue its execution but if this condition is not met then the algorithm will stop the algorithm computation. This equation is fulfilled when the sum of all individual elements related to each score assigned to each part number are equals to the risk matrix function.

$$Z\ (X_i) = \sum_{i=1}^{4} X_i \qquad (1)$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 September 2023 doi:10.20944/preprints202309.1812.v1

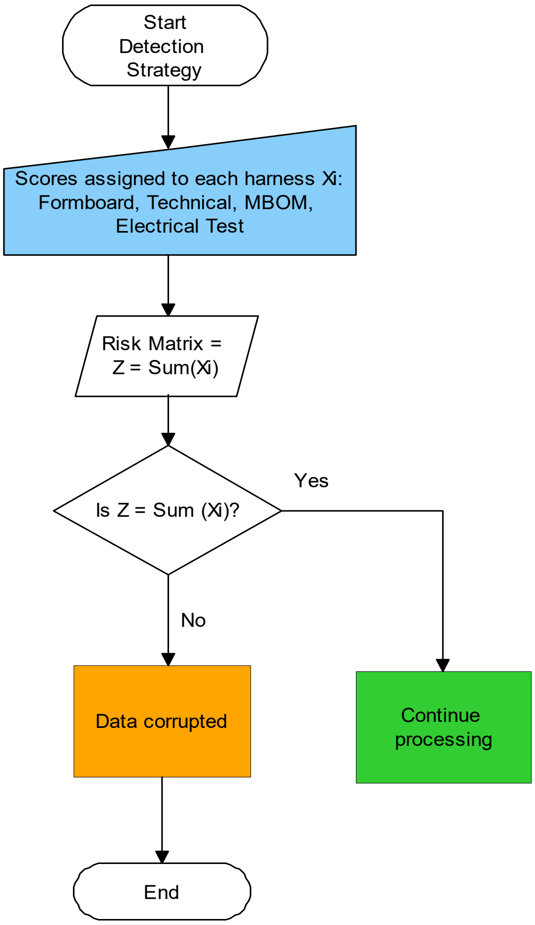
Start Detection Strategy
Scores assigned to each harness Xi: Formboard, Technical, MBOM, Electrical Test
Risk Matrix = Z = Sum(Xi)
Is Z = Sum (Xi)?
Yes
No
Data corrupted
Continue processing
End

Figure 6. Detection strategy to stop the algorithm computation after data has been successfully injected.

4. Conclusions

This study was performed at the electrical harness department in aerospace industry and it was based on the dataset of electrical harness manufactured and installed in a military aircraft C295. The experimental approach was based on impact of the data injection impacting on the main indicators used in the predictive algorithms which are necessary to develop manufacturing processes correctly. These indicators are, the risk matrix, the dendrogram and the confusion matrix as model outcomes performed through machine learning techniques within the artificial intelligent context. The first ones are well effectively assessing the risk prediction of error creation during the process of manufacturing engineering. The last one detects the rate between the real and predictive events in order to evaluate the performance of the algorithm use as detection mechanism. Overall, this study aimed to leverage machine learning techniques and artificial intelligence to enhance correct performance of the manufacturing processes of electrical harnesses in the aerospace industry.

The Table 1 summarizes the metrics outcomes related to the algorithm performance before and after the data has been injected. Before, the algorithm works correctly showing good excellent performance with very good metrics. These results suggest high level of reliability to make future predictions.

However, the findings after data injection show a negative impact on the performance of the algorithm. The data enrichment made the main metrics values and the performance of the algorithm decrease, being not possible to use it for future predictions. Consequently, the algorithm's effectiveness and reliability were compromised as a result of the modification data.

Security techniques are considered and developed to protect the algorithm and avoid malicious propagation to the outcomes. The strategy technique is used to secure the correct performance of the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 September 2023 doi:10.20944/preprints202309.1812.v1

algorithm. By prioritizing security measures, the integrity and reliability of the algorithm are kept, thereby preserving the accuracy and trustworthiness of its outputs. Synchronization of the outcomes are also guaranteed.

Based on the results of this study, it can be concluded that the comparisons identified between the algorithm performance before and after data injection provide valuable insights into the manufacturing process of electrical harnesses. The study highlights the importance of monitoring the metrics and to highlight the discrepancies found within the dataset in order to prevent occurrence of errors and enhance the reliability of the manufacturing processes.

The future work should be focus on analysis of the model variability from other aircraft with more extensive datasets. This situation will enable a more comprehensive assessment of the algorithm's performance across a wider range of aircraft types. Indeed, the use of diverse data will make this study further analyse of the robustness of the predictive model, enhancing its applicability and reliability in real-world scenarios.

Author Contributions: Conceptualization, methodology, software, analysis, writing J.B.H.; methodology review, visualization of section 3, supervision, M.M.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: We encourage all authors of articles published in MDPI journals to share their research data. The data presented in this study is available on request from the corresponding author. The data is not publicly available for confidentiality reasons.

Acknowledgments: The authors acknowledge the contributions to the electronic engineering department from the Seville University and to Airbus Poland for supporting this work with their research facilities and resources.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

CIA	Confidentiality, Availability and Integrity
ENISA	European Union Agency for Cybersecurity
APT	Advanced Persistent Threats
TP	True positives
TN	True negatives
FP	False positives
FN	False negatives

References

1. D. Catteddu and G. Hogben, "ABOUT ENISA," Cloud Computing: Benefits, risks and recommendations for information security, 2009.
2. M. Haseeb, H. I. Hussain, B. Ślusarczyk, and K. Jermsittiparsert, "Industry 4.0: A solution towards technology challenges of sustainable business performance," Social Sciences, vol. 8, no. 5, p. 154, 2019.
3. B. Jung, I. Han, and S. Lee, "Security threats to Internet: a Korean multi-industry investigation," Information & Management, vol. 38, no. 8, pp. 487-498, 2001.
4. Q. Su, H. Wang, C. Sun, B. Li, and J. Li, "Cyber-attacks against cyber-physical power systems security: State estimation, attacks reconstruction and defense strategy," Applied Mathematics and Computation, vol. 413, p. 126639, 2022.
5. J. Bautista-Hernández and M. Á. Martín-Prats, "A novel methodology to prevent failures in the manufacturing process using predictive algorithms through machine learning innovations for aerospace.," ed: [Manuscript in preparation], 2023.
6. P. Chen, L. Desmet, and C. Huygens, "A study on advanced persistent threats," in Communications and Multimedia Security: 15th IFIP TC 6/TC 11 International Conference, CMS 2014, Aveiro, Portugal, September 25-26, 2014. Proceedings 15, 2014: Springer, pp. 63-72.
7. J. Bautista-Hernández and M. Á. Martín-Prats, "Monte Carlo Simulation Applicable for Predictive Algorithm Analysis in Aerospace," in Doctoral Conference on Computing, Electrical and Industrial Systems, 2023: Springer, pp. 243-256.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 September 2023

doi:10.20944/preprints202309.1812.v1

8. P. Garcia-Teodoro, J. Diaz-Verdejo, G. Maciá-Fernández, and E. Vázquez, "Anomaly-based network intrusion detection: Techniques, systems and challenges," computers & security, vol. 28, no. 1-2, pp. 18-28, 2009.
9. X. Li, X. Liang, R. Lu, X. Shen, X. Lin, and H. Zhu, "Securing smart grid: cyber attacks, countermeasures, and challenges," IEEE Communications Magazine, vol. 50, no. 8, pp. 38-45, 2012.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.