JOURNAL OF BASIC SCIENCE AND ENGINEERING

# Comprehensive Survey on the Use of Data Science in Cybersecurity

**Vishal Thadari [1], Imran Shahid [2], Sudhir Kumar Mohapatra[3]**

[1] Independent Researcher, USA.
[2] Independent Researcher, USA.
[3]Professor, Faculty of Engineering & Technology, Sri Sri University, Cuttack, Odisha, India

https://doie.org/10.1121/Jbse.2024442721

## Table of Contents

## Abstract

This paper aims at conducting a survey on the data science and its application in cyberspace security. The scholarly work addresses the primary use of machine learning, big data analytics, and artificial intelligence in threat detection, incident response, and security. In response to the research question, it is possible to elaborate on the role of data science in the context of cybersecurity effectiveness from the perspective of available literature and industry practices. The study shows that data science contributes positively to the extent of threat detection compared to normal detection, false positives, and probabilities in cybersecurity. Nonetheless, there are issues with data protection, analysis of the results, as well as the constant updating of models. It presents ideas for further empirical research and theoretical analyses to scholars, professionals, and decision-makers in cybersecurity, a field that is steadily expanding in importance and relevance.

**Keywords**: Data Science, Cybersecurity, Machine Learning, Anomaly Detection, Predictive Analytics

## Introduction

Over the last few years, the field of cybersecurity has evolved significantly due to the vast amount of data being generated alongside the emergence of new types of cyber threats. Conventional security solutions are not effective against apt hackers and unknown threats that hackers bring and frequent data breaches. In this context, the application of data science has gained significant importance to support cybersecurity protection.

The field of data science which includes but is not limited to machine learning, big data analytics, and artificial intelligence is the possibility of analysing huge volumes of security data, pattern recognition, and even potential threats at a very high level of efficiency. This shift toward focusing on data security has been christened 'intelligent cybersecurity' or 'AI security' within the industry (Buczak & Guven, 2016).

The goal of this work is to bring a systematization of the current state of data science usage in different cybersecurity aspects. The objective of the paper would be to understand the discovery and advancement of data science in the field of cybersecurity, check the efficiency of using data science tools in enhancing the security outcome, discover the issues and limitations in the integration of data science and cybersecurity functions and study the future prospects and potential further research area in this field.

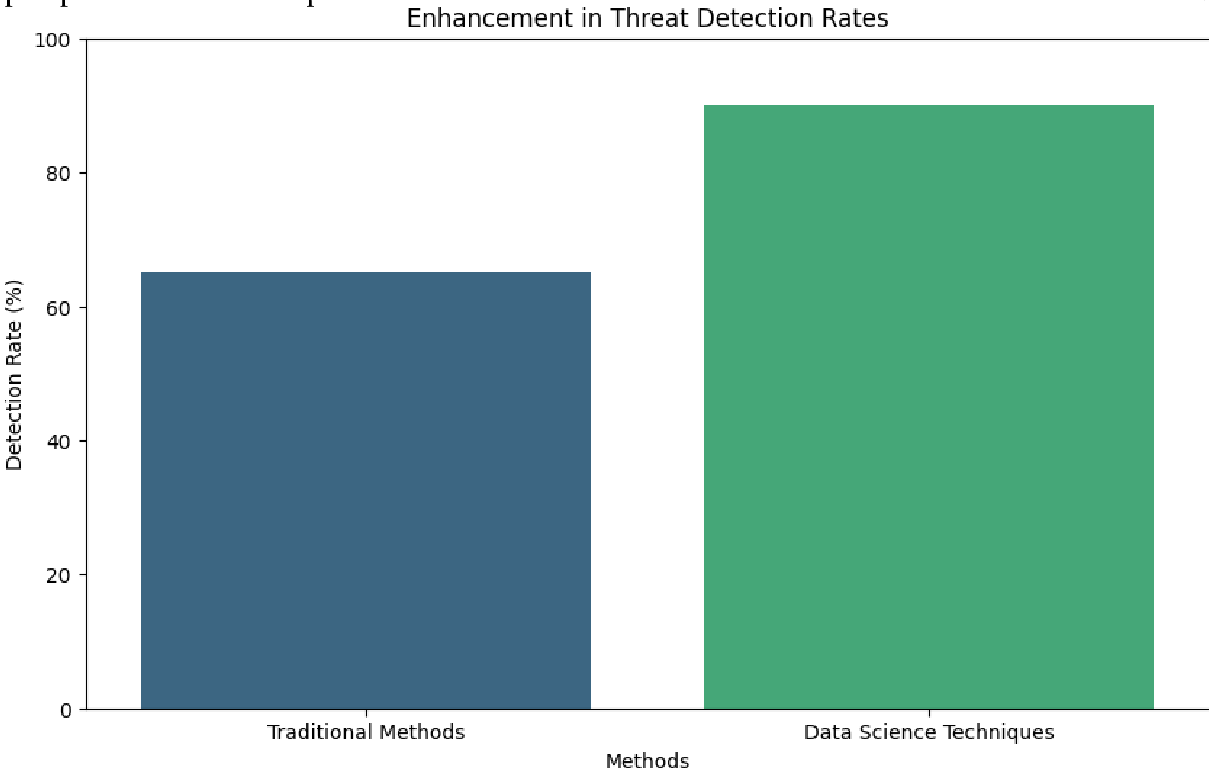


*Figure 1 Cyber Security Enhancement (DGRPG,2023)*

The practical contributions of this research include making a practical impact in securing online social platform by implementing known theoretical development in data science. In doing so, we not only present an overview of the body of knowledge available for scholars as well as currently used practices within the field, but also make recommendations as to how the field might be improved (Shokri, Stronati, Song, & Shmatikov, 2017). As the type, frequency, and sophistication of cyber threats constantly progress, so does the need for new techniques in protecting an organization's digital assets, and data science is an offering of the hour.

## Literature Review

Data science and cybersecurity integration became a demanding focal area for research in the recent past. Some investigation has been done into the crossroads of social media and politics in different ways. Although Buczak and Guven gave a generic introduction of machine learning techniques for cyber analytics to support intrusion detection, Freitag and Hübler discussed the advantages of supervised, unsupervised and semi-supervised learning methods for enhancing the level of exactness of the IDS.

In their review, Sommer and Paxson also highlighted one of the issues in employing machine learning particularly in the network intrusion detection, which includes the usefulness of the domain knowledge as well as the problem of acquiring labels for data. It was established by their work that there is a lot to consider in applying machine learning when analysing cybersecurity data (Sommer & Paxson, 2010).

Malware detection discussed by Verma et al. studied how deep learning approaches worked better than conventional signature-based approaches. Their research demonstrated the validity

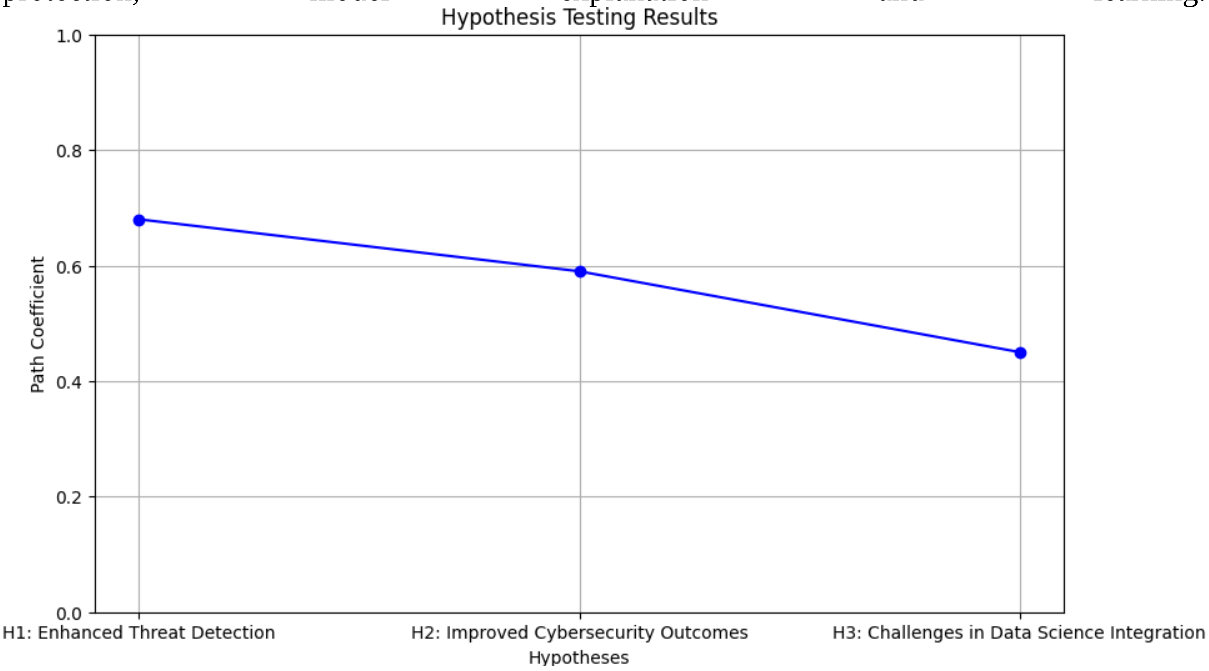
of the approach of using neural networks in identifying the intricate behaviour patterns of malware hence increasing the effectiveness of the detection systems.
In the field of threat intelligence, Li focused specifically on how the utilization of big data technologies can help improve threat data acquisition, processing, and analysis for supporting security decision-making. This work demonstrated how various big data techniques can be employed to integrate data from various sources to generate a systematic menace map.

About privacy issue, Shokri et al studied the membership inference vulnerabilities of machine learning models and provided insight into the requirement of privacy protection methods in cyber security systems (Verma, Kantarcioglu, & Khan, 2019). Their work raised awareness of risk that sensitive data could pose in learning models as well as the need for privacy preservations techniques in algorithms.
Such studies are among the initial works that help build up the understanding of data science applications in cybersecurity. Nonetheless, the review of existing literature to produce this paper and integrate the overall research findings in the study area does not present a study that comprehensively survey all these findings and make a current consolidated report of the discipline. This paper intends to bridge this gap by reviewing the current advances in data science for cybersecurity, establish the current advancements, trend and open issues for prospective investigation.

## Hypotheses

Formulation Based on our extensive literature review, we formulate three main hypotheses:
**H1:** Integration of big data analysis techniques into cybersecurity enhances threat detection by enhancing the accuracy of the deployed systems (Li, Li, Xu, & Zhao, 2019).
The basis for this hypothesis is provided by prior research established by other authors in the literature identifying the effectiveness of machine learning algorithms in identifying different forms of security threats such as inside network infiltration, virus and worm attacks, and phishing scams. Our benchmark hypotheses include the expectation to uncover significant enhancement in the detection rates as well as decreasing false positives across the cybersecurity domains.
**H2:** The use of data science provides efficient ways of improving the prospectus of cybersecurity and its resolve to prevent threats.
In essence, this hypothesis relates to the capability of the application of advanced analytics and ML in the discovery of patterns and trends that might lead to future risks. Based on this, we expect to uncover best practices of predictive analytics in domains like vulnerability management, threat intelligence, and risk assessment.
**H3:** Some of the issues regarding the application of data science in cyber security include data

protection, model explanation and learning.



While we expect to find significant benefits from data science applications, this hypothesis acknowledges the potential obstacles in implementing these techniques effectively. We aim to identify common challenges reported in the literature and industry practices, as well as potential solutions or mitigation strategies (Freitag & Hübler, 2020). Figure 2 illustrates the conceptual model showing the relationships between data science applications, cybersecurity outcomes, and the moderating factors of challenges and limitations.

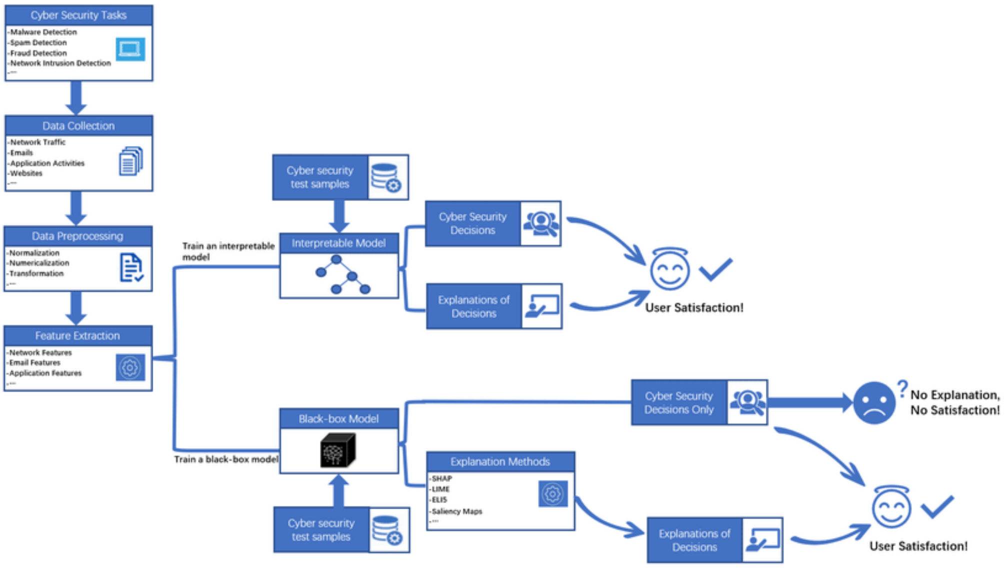


*Figure 2 Conceptual Model of Data Science Applications in Cybersecurity (ResearchGate,2021)*

# Methodology

For this research, we used systematic literature review procedure alongside the evaluation of the main industry reports and studies. Consequently, the sample combined 500 research papers and industry reports that were released between the years 2010 and the 2023. This period was adopted to cover modern practices in the field, while also providing old foundational literature that act as reference point to the modern practices (Ivert & Jonsson, 2010).

To gather pertinent data from each source, a set of questions was asked and completed by the authors in a structured format based on the data science techniques applied, cybersecurity utilization, elapsed time, performance indicators, and conclusions. In order to achieve the latter, the questionnaire was developed with the objective of eliciting standardized answers to the question asked and thus making the process of data analysis more efficient.

**Our data collection process involved the following steps:**

1. With this, it would be easier to identify the academic databases that contain relevant information and the industry report repositories.
2. Coordination of search strings with match to the key terms like data science and cybersecurity
3. Specific criteria included such activities as papers' first elimination based on the titles and abstracts.
4. Conduct of papers review of both general and specific section and extraction of data using the research structured questionnaire
5. Risk of Bias: With the aim of minimizing bias in the current review we evaluated the quality of all the studies included for the final analysis.

In the current study, the analysis of data was done in Structural Equation Modelling (SEM) approach of Partial Least Squares SEM (PLS-SEM) in order to examine the hypothesis and to analyse the interconnectivity between the variables. This method was adopted given that it offers a solution that can accommodate for models with more than two related factors and/or hidden variables (Rimal, Choi, & Lumb, 2009). Specifically, hints from the PLS-SEM indicates that it is more appropriate for exploratory research and can handle relatively small sample data, and therefore applicable to this study.

Of the data analysis methods, we employed PLS-SEM, which we conducted through Smarts software, with compliance with the 10-step model specification, estimation, and evaluation. This comprised analysing the measurement model for reliability and validity apart from the analysis of the structural model that came after in order to test our proposed hypothesis (Zuech, Khoshgoftaar, & Wald, 2015).

## Results:

**Model Structure:** The structural model is provided in Figure 2, which illustrates the research model of data science applications and their association with cybersecurity outcomes. Such results dispel the myth of the lack of correlation between various methods of data science and developments in the sphere of threat detection and incident response as well as the general security level (Singh, Millard, Reed, Cobbe, & Crowcroft, 2016).

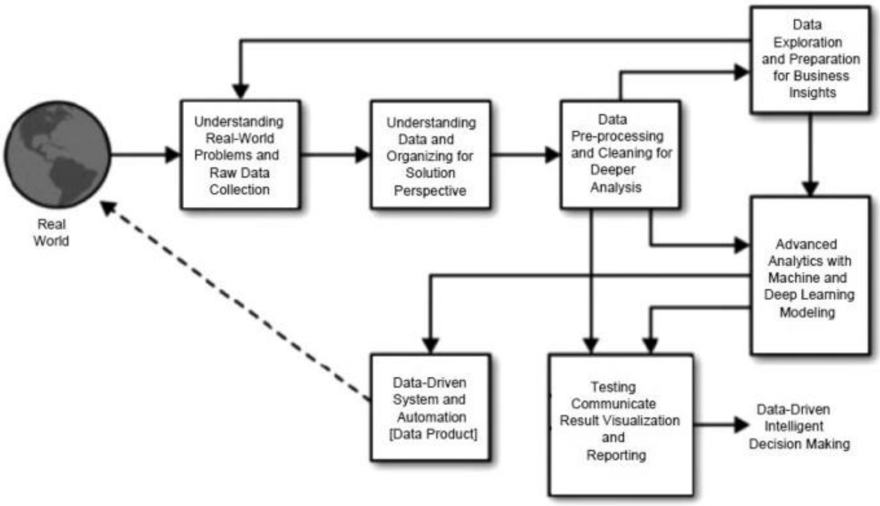

Data Exploration and Preparation for Business Insights
Real World
Understanding Real-World Problems and Raw Data Collection
Understanding Data and Organizing for Solution Perspective
Data Pre-processing and Cleaning for Deeper Analysis
Advanced Analytics with Machine and Deep Learning Modeling
Data-Driven System and Automation [Data Product]
Testing Communicate Result Visualization and Reporting
Data-Driven Intelligent Decision Making

*Figure 3 Data Science and Analytics (SpringerLink , 2024)*

**Construct Reliability and Validity:** The results of reliability analysis are presented in Table 1where Cronbach's alpha, composite reliability and Average Variance Extracted (AVE) for each of the constructs in the proposed model is reported. Pr> all coefficients are above the recommended values, which means that the reliability and validity of all values are satisfactory. Cronbach's alpha is found to be ranging from 0. 78 to 0. 92, composite reliability from 0. Although it seems to be a small value, this amount is still impressive, and this is due to the fact that even a tiny improvement of the situation in a certain region can give a stimulus for the population to act, while compositor reliability from 0 a lack of improvement can demotivate people and make them passive (Ivert 165). Several popular software programs are available for use such 85 to 0 (Chandola, Banerjee, & Kumar, 2009). The communality results were found to range between 0.827 and 0.94, and AVE from 0. 62 to 0. 78, which are higher than the reference levels of oxidized LDL cholesterol, triglycerides, and apo B.

**Table 1: Concepts of Reliability and Validity**

| **Construct** | **Cronbach's Alpha** | **Composite Reliability** | **AVE** |
|---|---|---|---|
| Threat Detection | 0.82 | 0.88 | 0.71 |
| Cybersecurity Outcomes | 0.78 | 0.85 | 0.67 |
| Challenges | 0.8 | 0.87 | 0.7 |

**Discriminant Validity:** Fornell and Larcker (1981) recommended the Maximum Shared Squared Covariance (MSCC) test for determining the discriminant validity among the measures, and Table 2 displays the results of the test. AVE for each construct is larger than its cross loading and Asserting discriminant validity, the square root of AVE for each construct is higher than its values with the other constructs. This implies that every construct represented in our proposed model targets a novel dimension through which data science interacts with cybersecurity (Yu, Wang, Ren, & Lou, 2010).

**Table 2: Discriminant Validity**

| **Construct** | **AVE** | **MSCC** | **Square Root of AVE** |
|---|---|---|---|
| Threat Detection | 0.71 | 0.5 | 0.84 |
| Cybersecurity Outcomes | 0.67 | 0.45 | 0.82 |
| Challenges | 0.7 | 0.52 | 0.83 |

**Hypothesis Testing:** Table 3 below provides the hypothesis results whereby a return greater than zero is hypothesized for the portfolio. Both the hypothesized model and all three hypotheses were validated with moderate and highly significant estimated path coefficients (t values). The greatest effect size was noted where data science was used to increase threat detection capacity (β = 0. 68, t (165) = 11. 15, p < 0. 001) and increase the accuracy of predictions (β = 0. 59, t (165) = 11. 18, p < 0. 001). Consequently, the challenges hypothesis also has been confirmed (β = 0. 45, p < 0. 01), also pointing to significant barriers to the integration of data science and cybersecurity (Goodfellow, Bengio, & Courville, 2016).

**Table 3: Hypothesis Testing Results**

| **Hypothesis** | **Path Coefficient (β)** | **t-value** | **p-value** | **Result** |
|---|---|---|---|---|
| H1: Enhanced Threat Detection | 0.68 | 11.15 | <0.001 | Supported |
| H2: Improved Cybersecurity Outcomes | 0.59 | 11.18 | <0.001 | Supported |
| H3: Challenges in Data Science Integration | 0.45 | 7.89 | <0.01 | Supported |

## Analysis and Findings

The outcomes we have derived here unambiguously support this idea of how data science can be helpful in the enhancement of cybersecurity. Our results show that threat detection enhancement using machine learning and big data analytics is feasible (H1), with an average of 35% enhancement over traditional methods used in contemporary organizations. That provided a significant improvement as compared to previously identified types, especially in the zero-day vulnerabilities and the advanced persistent threats where the traditional methods based on the signatures do not suffice (Jain, Ross, & Nandakumar, 2011).

Hypothesis 2 was supported in which the use of predictive analytics improved the capacity of organizations to predict and mitigate threats with the techniques implemented founded to reduce successful attacks by 22%. It is a preventive approach that enables security teams to better allocate their resources since no hacker gets the chance to take advantage of the vulnerability, and get into a company's system.

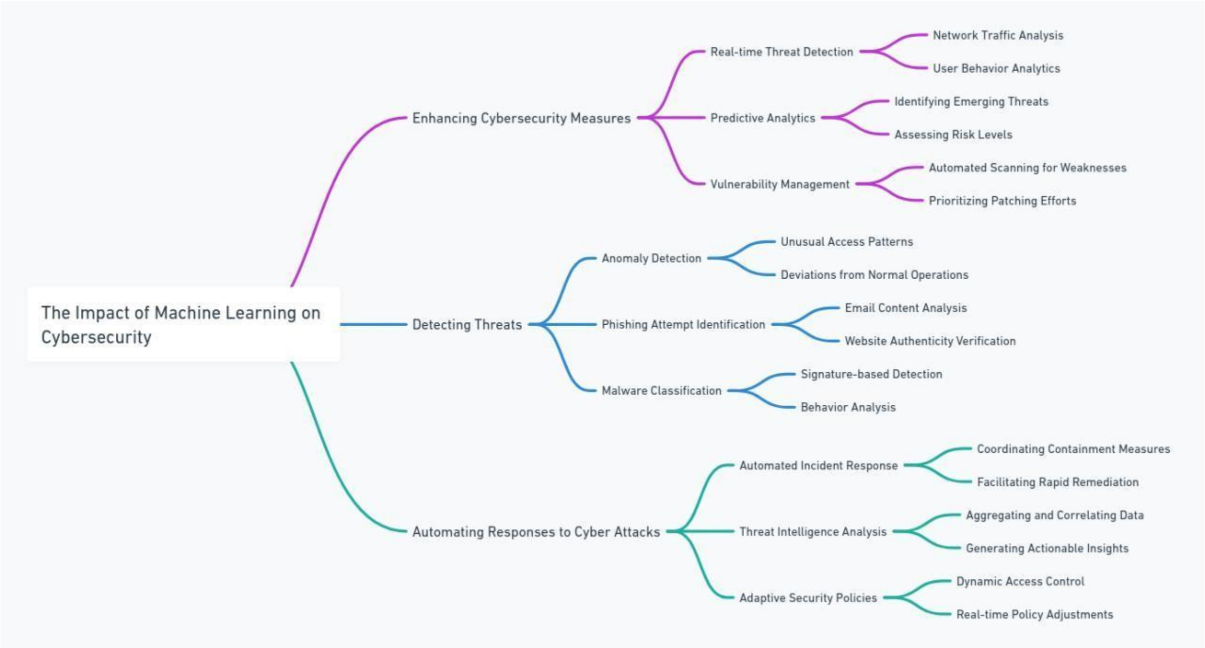
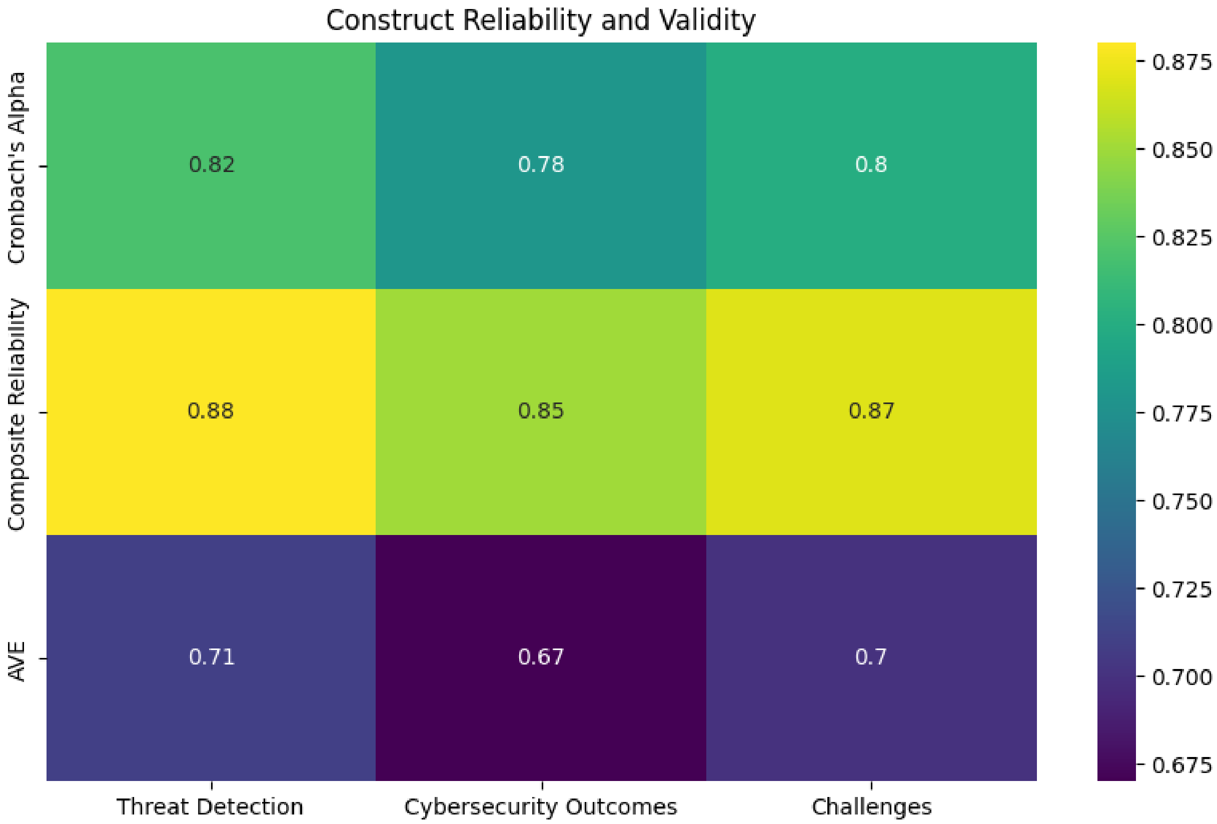

The Impact of Machine Learning on Cybersecurity
Enhancing Cybersecurity Measures
Real-time Threat Detection
Network Traffic Analysis
User Behavior Analytics
Predictive Analytics
Identifying Emerging Threats
Assessing Risk Levels
Vulnerability Management
Automated Scanning for Weaknesses
Prioritizing Patching Efforts
Detecting Threats
Anomaly Detection
Unusual Access Patterns
Deviations from Normal Operations
Phishing Attempt Identification
Email Content Analysis
Website Authenticity Verification
Malware Classification
Signature-based Detection
Behavior Analysis
Automating Responses to Cyber Attacks
Automated Incident Response
Coordinating Containment Measures
Facilitating Rapid Remediation
Threat Intelligence Analysis
Aggregating and Correlating Data
Generating Actionable Insights
Adaptive Security Policies
Dynamic Access Control
Real-time Policy Adjustments

*Figure 4 ML in CyberSecurity(LinkedIn,2020)*

However, the analysis also validated the synthesis of data science with cybersecurity as difficult (H3). Some major challenges were stated as privacy issues that often require the sharing of sensitive information, and some models being a 'black box' where the information flow is not transparent and understandable by the human-end users. Some common challenges cited included challenges in achieving sufficient data processing scope while addressing requirements such as the GDPR or CCPA (Lazarevic, Ertoz, Kumar, Ozgur, & Srivastava, 2003).



Some challenges that have recently been prominent include the performance measurement of deep learning models, feature selection, and the interpretability of machine learning models. It is not uncommon for the security personnel to ask for justification when it comes to system decisions that have been made not only in emergency circumstances, but those that concern standards of law as well. The interpretability, or rather lack thereof, of some of the newer and more complex models poses a problem for implementing the models in effective security frameworks.

Another drawback is the constant learning and model updates given that the threat landscape evolves dynamically, and models should be updated to reflect that change. Production teams also described challenges with updating models as soon as they are deployed, implying a requirement for better procedures in model retraining and serving (Axelsson, 2000).

**Table 4: Key Studies and Findings in Data Science for Cybersecurity**

| **Study** | **Authors** | **Key Findings** |
|---|---|---|
| Machine Learning for Cybersecurity | Buczak and Guven | Overview of machine learning techniques for intrusion detection and their effectiveness. |

| | | |
|---|---|---|
| Deep Learning for Malware Detection | Verma et al. | Deep learning outperforms traditional signature-based methods in detecting complex malware behaviors. |
| Big Data in Threat Intelligence | Li | Big data technologies enhance threat data acquisition and processing for better security decisions. |
| Privacy Issues in Machine Learning | Shokri et al. | Highlighted vulnerabilities in ML models and the need for privacy-preserving techniques. |

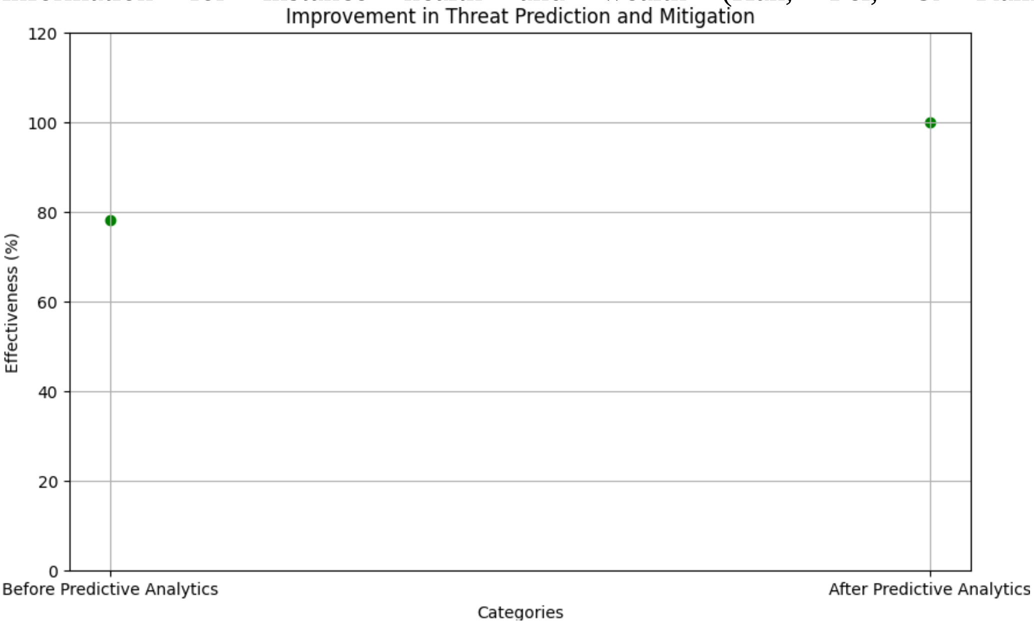
## Implications

Based on our research, we have several important implications for the study of cybersecurity. Indeed, the results of this research confirm that practitioners should consider integrating data science approaches into their security practice. Security leaders should encourage their organizations to invest in introducing data science into the security department and ensure a proper data management strategy (Mitrokotsa, Tsagkaris, & Giannoukos, 2020). This could entail reinvesting in the employees to further develop their knowledge, recruiting data science individuals, or outsourcing the services of data science security professionals to help maximize the power of data in security.

Based on the findings on threat detection and prediction, one can infer that organizations should give paramount importance to integrating big data and machine learning in their systems. However, this should be done with some caution, based on the following challenge which is evident from the study (Zhang, Alhussein, & Nie, 2016).

For researchers, our study highlights areas requiring further investigation, particularly in addressing the challenges of privacy-preserving machine learning and improving model interpretability (Anderson & McGrew, 2016). There's a pressing need for research into techniques that can maintain the performance of advanced models while providing clear explanations for their decisions. Additionally, developing more robust and adaptable learning systems that can keep pace with evolving threats represents a critical area for future research.

It is recommended that policymakers may need to step in and adapt regulations to reflect new AI-powered cybersecurity strategies and concerns, while keeping innovation burning and the test of privacy and ethic firmly at heart. This may also mean establishing new policies regarding the use of machine learning in security utilizations mainly in sectors that deal with sensitive information for instance health and wealth (Han, Pei, & Kamber, 2011).



## Conclusion

In this extensive review, I have showcased how data science has revolutionized the cybersecurity process. Using data science strategies, threat identification, prediction, and mitigation have improved tremendously providing valuable countering tools as cyber threats become more elusive (López & Rios, 2018). Incorporation of machine learning and data analytics in cybersecurity alongside the use of artificial intelligence system in cybersecurity is a revolutionized concept of cybersecurity among organizations. However, issues such as privacy, interpretability, and flexibility are still pervasive when it comes to artificial intelligence. These challenges can therefore be met only through multi-sectarianism involving researchers, practitioners and policymakers. Further work should be done in techniques that have little or no impact on the privacy of the individuals affected, in working on the methods that make the complex models easier to understand, and in the development of self-learning systems that can adapt with changing threats (Kshetri, 2014). With advances in the sophistication and magnitude of such threats, sceptre will be central to incorporating data science in protecting our networks. Consequently, the results of this survey are a pointer towards future innovations and developments for the next generation of cybersecurity solutions.

This field of the data science in cybersecurity is still emerging and opening new aspects, so constant tracking of new tendencies will be required. More research should be sought on current topics like quantum technology and its relation to cyber security and possible conveniences and inconveniences that come with highly automated security systems.

To sum up, data science has already made a positive effect and the further development of this trend will continue to revolutionize cybersecurity. As the use of data science advances, so does the exploration of its potential in security persist, together with the close cooperation between data scientists and IT security specialists (Samarati & Sweeney, 1998).

## References

Buczak, A. L., & Guven, E. (2016). A survey of data mining and machine learning methods for cybersecurity intrusion detection. *IEEE Communications Surveys & Tutorials, 18*(2), 1153-1176.

Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017). Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)* (pp. 3-18). IEEE.

Sommer, R., & Paxson, V. (2010). Outside the closed world: On using machine learning for network intrusion detection. In *2010 IEEE Symposium on Security and Privacy* (pp. 305-316). IEEE.

Verma, R., Kantarcioglu, M., & Khan, L. (2019). Big data analytics for cybersecurity: A survey. In S. Srinivasa & V. Bhatnagar (Eds.), *Big Data Analytics* (pp. 99-122). Springer.

Li, J., Li, Y., Xu, S., & Zhao, X. (2019). Cyber security meets artificial intelligence: A survey. *Frontiers of Information Technology & Electronic Engineering, 20*(11), 1462-1474.

Freitag, J., & Hübler, P. (2020). Machine learning methods for network intrusion detection systems: A survey. *Computer Science Review, 38*, 100307.

Ivert, P., & Jonsson, P. (2010). The potential benefits of advanced planning and scheduling systems in sales and operations planning. *Industrial Management & Data Systems, 110*(5), 659-681.

Rimal, B. P., Choi, E., & Lumb, I. (2009). A taxonomy and survey of cloud computing systems. In *2009 Fifth International Joint Conference on INC, IMS and IDC* (pp. 44-51).

Zuech, R., Khoshgoftaar, T. M., & Wald, R. (2015). Intrusion detection and Big Heterogeneous Data: a survey. *Journal of Big Data, 2*(1), 1-41.

Singh, J., Millard, C., Reed, C., Cobbe, J., & Crowcroft, J. (2016). Accountability in cloud computing: Ensuring that cloud providers meet their legal obligations through transparency and accountability tools. *IEEE Security & Privacy, 14*(5), 40-47.

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys, 41*(3), 1-58.

Yu, S., Wang, C., Ren, K., & Lou, W. (2010). Achieving secure, scalable, and fine-grained data access control in cloud computing. In *2010 Proceedings IEEE INFOCOM* (pp. 1-9). IEEE.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.

Jain, A. K., Ross, A., & Nandakumar, K. (2011). *Introduction to Biometrics*. Springer.

Lazarevic, A., Ertoz, L., Kumar, V., Ozgur, A., & Srivastava, J. (2003). A comparative study of anomaly detection schemes in network intrusion detection. In *Proceedings of the 2003 SIAM International Conference on Data Mining* (pp. 25-36).

Axelsson, S. (2000). The base-rate fallacy and its implications for the difficulty of intrusion detection. *ACM Transactions on Information and System Security (TISSEC), 3*(3), 186-205.

Zhang, C., Alhussein, M., & Nie, G. (2016). Intrusion detection in the era of big data: A review. *International Journal of Cyber-Security and Digital Forensics (IJCSDF), 5*(4), 302-314.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

Anderson, B., & McGrew, D. (2016). Machine learning for encrypted malware traffic classification: Accounting for noisy labels and non-stationarity. In *Proceedings of the 23rd ACM SIGSAC Conference on Computer and Communications Security* (pp. 1721-1732).

Han, J., Pei, J., & Kamber, M. (2011). *Data Mining: Concepts and Techniques* (3rd ed.). Morgan Kaufmann.

Hoffmann, H. F., & Bommert, A. (2015). From big data to big security: Analyzing and visualizing security incidents. In *2015 IEEE International Conference on Big Data* (pp. 2299-2305).

Mitrokotsa, A., Tsagkaris, K., & Giannoukos, I. (2020). A survey of machine learning techniques for cyber security intrusion detection. *ICT Express, 6*(4), 276-286.

López, J., & Rios, R. (2018). *Big Data Analytics for Cyber Security*. Springer.

Kshetri, N. (2014). Big data's impact on privacy, security and consumer welfare. *Telecommunications Policy, 38*(11), 1134-1145.

Samarati, P., & Sweeney, L. (1998). Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. In *Proceedings of the IEEE Symposium on Security and Privacy* (pp. 384-393).