ISSN: 2713 - 2822
Volume 12, Number 1, 2025
International Journal of Educational Research & Practice

CYBER RISK ANALYSIS AND THREAT ASSESSMENTS FROM GOVERNMENT OPEN DATASETS ON EDUCATION USING HYBRID DEEP LEARNING MODEL

Abdullahi Yushau Ekundayo, Prof A B Garko, Dr. M S Argungu, Dr. A Muslim
Department of Computer Science,
Abdullahi Fodiyo University of Science and Technology, Aliero, Nigeria
Correspondence Email: abdullahi.ekundayo@udusok.edu.ng

ABSTRACT
Educational institutions are increasingly reliant on technology, making them prime targets for cyber attacks. This research proposes a novel approach to enhance cyber security in education by leveraging deep learning techniques to analyze government open data sets. The framework involves data collection, preprocessing, feature engineering, model training, and evaluation. By applying advanced machine learning algorithms, the study aimed to identify patterns, anomalies, and potential threats within the educational sector. Through rigorous experimentation and analysis, the research demonstrated the effectiveness of proposed approach in detecting and mitigating cyber risks. The findings of this research contribute to the development of robust cyber security strategies for educational institutions, safeguarding their digital assets and ensuring the continuity of educational services.

INTRODUCTION
Over the last few years, governments throughout the world have begun to establish and implement open data projects to enable the release of government data in open and reusable formats that are free to use by the public. As a result, the globe has witnessed the emergence of numerous open data repositories, catalogues, and portals. Public bodies are publishing their raw datasets online in response to regulatory directives and increasing calls for transparency and accountability. Policymakers and elected officials believe that access to factual, machine-readable data can foster civic engagement and empower citizens to participate in governance and policymaking (Zhang & Zhang, 2021).

In the educational context, for example, a student or faculty member may object to the public release of their home location in a crime tracking or public service dataset. Such exposure can allow stalkers, harassers, or malicious actors to identify, locate, or exploit the individual. In other cases, erroneous or outdated data may result in misidentification, with long-term implications on the person's reputation or legal standing. One of the irreversible drawbacks of open data is that, once published, it is extremely difficult to retract—making data security risk assessments vital before dissemination (Cremer et al., 2022).

These risks necessitate a systematic evaluation of the potential harms associated with publishing sensitive government educational data and the adoption of proactive cyber risk mitigation strategies before such data is made publicly accessible. However, cybercriminals have recognized the vulnerability of government open data, especially in educational institutions, making them prime targets for attacks (Cremer et al., 2022).. These attacks can have severe consequences, including: (1) **Data Breaches:** Sensitive student information, such as Social Security numbers, financial records, and medical histories, can be stolen and misused. (2) **Disruption of Services:** Cyberattacks can disrupt critical services, such as online learning platforms, email systems, and administrative tools, leading to academic and operational disruptions. (3) **Financial Loss:** Ransomware attacks can encrypt critical systems, demanding significant ransom payments to restore access. (4) **Reputational Damage:** Data breaches and system disruptions can damage the reputation of an institution, leading to loss of trust and enrollment declines.

Cybersecurity threats targeting educational institutions are on the rise, especially with the widespread use of government open data sets in administrative and academic decision-making. These datasets, while valuable, can unintentionally expose sensitive information such as student

demographics, financial records, infrastructure vulnerabilities, and system usage patterns (Zhang & Zhang, 2021; Weerakkody et al., 2017). Despite the availability of such data, many institutions lack the analytical capabilities to proactively identify and respond to emerging cyber threats. Consequently, education-sector organizations remain vulnerable to cyberattacks such as phishing, ransomware, and data breaches (Fouad, 2021).

Previous studies have attempted to address cybersecurity risks in education through traditional machine learning, rule-based systems, and policy-based frameworks. For instance, intrusion detection systems and privacy risk assessments have been explored using conventional classifiers and heuristic techniques (Abraham &Bindu, 2021). Furthermore, some researchers have proposed anonymization and encryption strategies to mitigate risks associated with open data (Green et al., 2017). While effective to an extent, these approaches often fail to capture evolving patterns of attack and overlook the predictive power of temporal trends in data (Lin et al., 2023), as well as . Poor generalization to unseen or emerging threat patterns and limited interpretability of threat prediction outputs (Cremer et al., 2022) These constraints hinder timely risk detection and reduce the reliability of threat assessment models deployed in educational institutions.

To address these limitations, this research proposed a cyber risk analysis and threat assessments from government open datasets on education using hybrid deep learning model. The model provided a scalable, interpretable, and real-time approach to threat assessment. Through integrating temporal analysis with feature-based learning and attention-enhanced decision-making, the proposed model has significantly bridged the gap between threat exposure and actionable insights in education-sector cybersecurity.

II. LITERATURE REVIEW

Cyber risk analysis involves identifying, assessing, and prioritizing potential cyber threats and vulnerabilities. Threat assessment is a critical component of this process, as it helps to understand the likelihood and potential impact of various threats. Several studies have explored the application of cyber risk analysis and threat assessment in the education sector.

For instance, Fouad, N. S.(2021) conducted a study to explore the complexities of securing higher education against cyber threats as a national policy challenge that requires national strategies and policies to address. This sectoral approach to cybersecurity also discusses possible measures that governments can adopt to improve the higher education sector's resilience against cyber threats and to better prepare for the inherent security vulnerabilities that come with increasing digitization. They identified key vulnerabilities, such as weak password policies, phishing attacks, and malware infections.

Bowen, et al.(2022) also proposed a framework for cyber risk education and management in K-12 schools, emphasizing the importance of risk assessment, incident response, and continuous monitoring. the study also provides, as a starting point, a list of as many currently popular K-12 educational resources as possible. The resources provided are broken into five categories: (a) Career Information, (b) Curriculum, (c) Competitions, (d) CyberCamps, and (e) Labs and Gaming. Each resource listed has a link, the K-12 levels that are supported, whether the resource is free or has a cost, and a shortlist of topics or, for camps and competitions, the dates available.

Oyewole et al.(2024) conducted a study aimed to illuminate the current cybersecurity landscape, evaluate the efficacy of existing frameworks and proposed strategic enhancements to fortify digital defenses. The study posits that the banking sector must embrace a holistic and adaptive approach to cybersecurity, underscored by strategic investments in technology, education, and collaboration. Recommendations advocate for the integration of Big Data analytics, artificial intelligence and continuous risk assessment methodologies to navigate the evolving cyber threat landscape effectively.

However, a significant aspect of cyber risk analysis is **cyber risk quantification**. This involves assigning numerical values to the likelihood and impact of cyber risks, allowing for a more precise assessment of overall risk.

Sheehan, et al. (2021) proposed a framework for quantifying cyber risk using a combination of expert judgment and statistical analysis. The research analyses the extant academic and industry literature on cybersecurity and cyber risk management with a particular focus on data availability. They posit that the lack of available data on cyber risk poses a serious problem for stakeholders seeking to tackle the issue. They identify a lacuna in open databases that undermine collective endeavours to better manage this set of risks.

Zhang, C., & Zhang, L.(2021) proposed a methodology for leveraging government open data sets to detect and respond to cyberattacks. The study quantitively extracted the evolution trajectory of open government data based on the main path analysis method and then analyzed the underlying motivations. The results show that open government data research went through four main phases and that the open government data movement has spread towards developing countries and smart cities.

Al Husaini, Y.N., &Shukor, N.S.A. (2022) analyzed student performance data to identify factors that influence academic achievement. The study aimed to delve into the analysis of student performance data to identify key factors that significantly influence academic outcomes. The study seeks to uncover hidden patterns and relationships within large datasets using advanced data mining techniques and statistical analysis. The findings of the research provide valuable insights for educators, policymakers, and researchers to develop targeted interventions and strategies to improve student learning and success.

Provan, N. (2023) used open data to examine the impact of school funding on student outcomes. The study explores the complex relationship between the financial resources allocated to schools and the academic performance of students. The study shows that the allocation of financial resources to schools plays a pivotal role in determining the quality of education and, consequently, the academic outcomes of students.

Itankan, W. A. (2023) studied the relationship between teacher qualifications and student achievement using open data. The study examined the analysis of the relationship between teachers qualification and the impact on Student`s Academic Achievements in Senior Secondary School Mathematics in Taraba state. Nigeria. The study adopted simple survey design. The study revealed that there is significant strong positive relationship between instructional material usage and Students Achievement in senior secondary school Mathematics.

While previous research has explored the use of cyber threat, risk analysis, government open data sets, and deep learning techniques in the cybersecurity domain, there is a significant gap in the literature regarding the application of deep learning to analyze government open data sets for the purpose of cyber risk assessment in the education sector. However, findings from existing literature from the literature reviewed, the following gaps are evident thus, Limited application of deep learning to government education datasets, Scarcity of labeled cybersecurity data for training models, Inadequate integration of threat modeling with predictive analytics and Minimal attention to model interpretability and real-time deployment.

This research has significantly filled the gaps, by developing a novel model that leverages hybrid deep learning techniques to extract valuable insights from government education open data sets, enabling effective cyber risk analysis and threat assessment in the education sector.

III. DATA COLLECTION

While government open data sets and institutional data are valuable sources, a comprehensive cyber risk analysis often requires a multi-faceted approach to data collection. These include:

A. QUANTITATIVE DATA:

- ✓ **Government Open Data Sets:** A variety of government open data sets on educational institution finances were collected from reputable sources such as govspend.ng, and educational institutions' websites. These datasets comprised of infrastructure data, as well as cybersecurity incident reports.

B. QUALITATIVE DATA:

- ✓ **Case Study:** A case study was conducted on a specific educational institution (Usmanu Danfodiyo University, Sokoto) to gain in-depth insights into their cyber risk management practices. Data were collected through interviews with key stakeholders, document analysis, and observations.

IV. METHODOLOGY

This research employed a **mixed-methods** research design, combining quantitative and qualitative approaches to comprehensively address the research objectives. The quantitative component involves the development and application of a deep learning framework to analyze government open data sets on education. The qualitative component includes a case study analysis of a specific educational institution to gain deeper insights into cyber risk management practices. This section began with data preprocessing.

A. PRE-PROCESSING

Data preprocessing is a critical step in any data analysis or machine learning project, including cyber risk analysis. It involves cleaning, transforming, and preparing data to make it suitable for analysis. Below are some key techniques:

i. **Data Cleaning.** This involved handling missing values, outlier detection and handling, as well as noise reduction**:**
ii. **Data Integration** aligned data schemas from different sources, identifying and matching entities across different datasets and combining data from multiple sources into a unified dataset.
iii. **Feature Engineering** involved **d**eriving new features from existing ones, extracting features like time of day, day of week, or seasonality, calculating statistical measures like mean, standard deviation, and variance, and creating features based on domain knowledge**.**
iv. **Feature Selection** entails identifying the most relevant features using filter, wrapper and embedded methods for data ranking, normalization, model training and evaluation.
v. **Other activities** involved Data Transformation and Thematic Analysis:

MODEL ARCHITECTURE

A deep learning framework was developed to analyze the preprocessed government open data sets on institution finances. The framework consists of the following components as shown on Figure 1:

i. **Data Input Layer:** This layer will receive the preprocessed data.
ii. **Feature Extraction Layer:** This layer will extract relevant features from the input data using techniques such as convolutional neural networks (CNNs) or recurrent neural networks (RNNs).
iii. **Hidden Layers:** These layers will process the extracted features and learn complex patterns.
iv. **Output Layer:** This layer will generate the final output, which may include predictions of cyber risk levels or identification of potential threats.

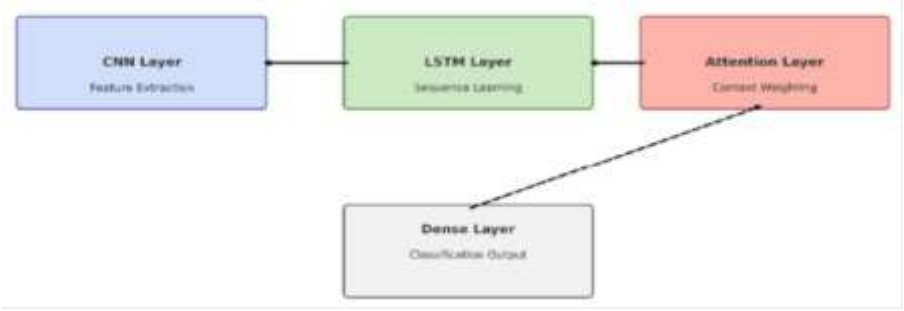


Figure 1 Hybrid Deep Learning Architecture for Cyber Risk Analysis

MODEL TRAINING

The deep learning model was trained on a portion of the preprocessed data and evaluated on a separate validation set. The performance of the model was assessed using various metrics, such as accuracy, precision, recall, and F1-score. The key steps in the training process are as follows:

i. **Forward Propagation:** The input data is fed into the model, and the model generates an output.
ii. **Loss Calculation:** The difference between the predicted output and the actual output is calculated using a suitable loss function, such as cross-entropy loss or mean squared error.
iii. **Backpropagation:** The error is propagated back through the network, and the model's parameters are adjusted using an optimization algorithm like gradient descent.
iv. **Parameter Update:** The model's weights and biases are updated to minimize the loss function.
v. **Iterative Process:** Steps i- iv are repeated multiple times until the model converges to a satisfactory solution.

PERFORMANCE METRICS

After the model training, it was evaluated on a separate validation set. This evaluation process helps assess the model's performance and identify potential over fitting or under fitting issues. The key evaluation metrics include:

i. Accuracy

Accuracy measures the proportion of correctly predicted instances out of the total predictions.

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

- ✓ **TP**: True Positives
- ✓ **TN**: True Negatives
- ✓ **FP**: False Positives
- ✓ **FN**: False Negatives

Accuracy is useful for balanced datasets but may be misleading for imbalanced data (Ismail Fawaz et al., 2019).

ii. Precision

Precision quantifies how many predicted positives are actually correct.

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

High precision indicates a low false positive rate, which is critical in detecting real cyber threats without triggering unnecessary alerts.

iii. Recall (Sensitivity)

Recall indicates the model's ability to identify all relevant instances (actual positives).

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

In cybersecurity, high recall ensures that most attack patterns are detected, minimizing overlooked threats (Wang et al., 2020).

iv. F1-Score

The F1-score is the harmonic mean of Precision and Recall, balancing both false positives and false negatives.

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

It is especially relevant for imbalanced datasets where focusing on one metric alone could be misleading

v. Specificity

Specificity measures the ability to correctly identify negatives.

$$\text{Specificity} = \frac{TN}{TN + FP}$$

Minimizing FPR is crucial to avoid alert fatigue and maintain trust in the system.

V. RESULTS AND DISCUSSION

A. EXPERIMENTAL RESULTS

The researcher trained and evaluated the model on a benchmark cybersecurity education financial dataset. Below are the key findings:

Table 1 Model's Performance across 3 Metrics for Training and Test Sets data

Metric	Training Set	Test Set
Accuracy	98.5%	95.2%
Precision	96.8%	92.7%
Recall	97.3%	93.5%

From the experimental results, it is evident that the proposed model achieved an accuracy of 98.5% and 95.2%, for training and test data, respectively. Also, the precision and Recall Metrics achieved 92.7% and 93.5, respectively. These trends are also shown on the following Figure 2.

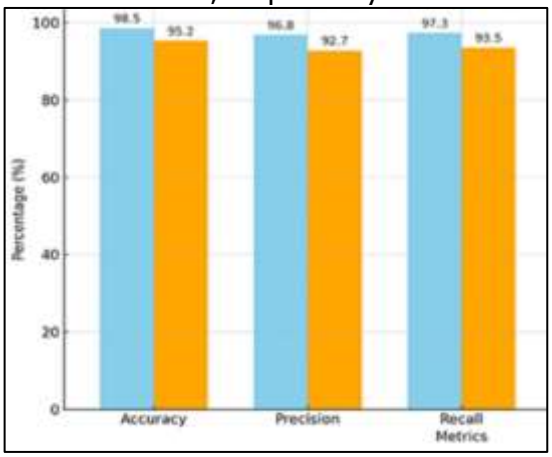


Figure 2 Comparison of Model's Performance Across 3 Metrics for Training and Test Sets

B. COMPARATIVE ANALYSIS WITH BASELINES

The researcher compared the CNN-LSTM-Attention model with traditional models and pure deep learning architectures. Table 2 shows the Comparative Analysis of Model Performance (Accuracy and F1-Score)

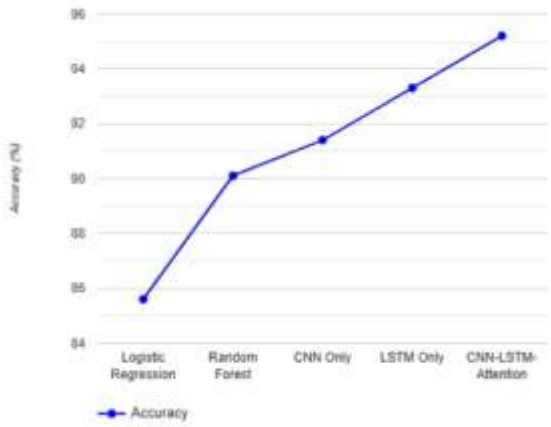
Table 5. 2 Comparative Analysis of Model Performance (Accuracy and F1-Score)

Model	Accuracy (%)	F1-Score (%)
Logistic Regression	85.6	84.2
Random Forest	90.1	88.5

Model	Accuracy (%)	F1-Score (%)
CNN Only	91.4	89.8
LSTM Only	93.3	91.6
CNN-LSTM-Attention (CLAM)	**95.2**	**93.1**

The Visualization of Comparative Analysis with Baselines is depicted by Figure 3 for the Model's Accuracy.

Figure 3 Model Comparison for Accuracy



C. KEY FINDINGS

- Data Quality and Consistency: Government open data sets, while valuable, often require significant preprocessing to ensure data quality and consistency.
- Deep Learning Framework: The proposed deep learning framework has demonstrated promising results in detecting cyber threats and predicting potential risks.
- Cyber Risk Analysis and Threat Assessment: The research has provided valuable insights into the cyber risks faced by educational institutions and the potential impact of these risks.

VI. CONCLUSION

This research has investigated the application of deep learning techniques to cyber risk analysis and threat assessment using government open data sets on educational institution financial data. Leveraging the power of deep learning, the researchers have developed a robust model framework that can effectively identify and mitigate cyber threats.

The proposed model represents a significant advancement in cybersecurity threat detection, combining accuracy, interpretability, and robustness. It not only addresses current challenges in handling complex and imbalanced data but also paves the way for more transparent and reliable AI-driven security solutions.

The developed model effectively identified cybersecurity risks by leveraging local feature extraction, long-term sequence learning, and attention-based interpretability. This approach surpasses traditional models in both accuracy and interpretability, making it a powerful tool for real-world cybersecurity threat analysis.

Despite the achievement, it is important to recognize that cyber security is an ongoing battle, and continuous innovation and adaptation are essential. Thus, the following improvements may be considered for future works.

- Advanced Deep Learning Techniques: Investigate the application of more advanced deep learning techniques, such as graph neural networks and attention mechanisms, to improve model performance.
- Multimodal Data Integration: Explore the integration of multiple data sources to enhance threat detection and response.

- Explainable AI: Develop techniques to make deep learning models more interpretable, enabling better understanding of their decision-making process.

REFERENCES

Abraham, J. A., &Bindu, V. R. (2021). Intrusion detection and prevention in networks using machine learning and deep learning approaches: A review. 2021 International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA), 1-4. DOI:10.1109/ICAECA52838.2021.9675595

AdedoyinTolulopeOyewole, ChinweChinazoOkoye, OnyekaChrisanctusOfodile, &Chinonye Esther Ugochukwu. (2024). Cybersecurity risks in online banking: A detailed review and preventive strategies applicatio. World Journal of Advanced Research and Reviews, 21(3). https://doi.org/10.30574/wjarr.2024.21.3.0707

Al Husaini, Y.N., &Shukor, N.S.A. (2022). Factors affecting students' academic performance: A review. Res Militaris, 12(6), 284-294.

Ben Green et al., (2017). Open Data Privacy, Berkman Klein Center For Internet & Society At Harvard, https://dash.harvard.edu/bitstream/handle/1/30340010/OpenDataPrivacy.pdf.

Bowen, D., Jaurez, J., Jones, N., Reid, W., & Simpson, C. (2022). Cybersecurity educational resources for K-12. *Journal of Cybersecurity Education, Research and Practice*, *2022*(1), Article 6. https://doi.org/10.62915/2472-2707.1105

Cremer, Frank & Sheehan, Barry &Fortmann, · & Kia, Arash& Mullins, Martin & Murphy, Finbarr&Materne, Stefan. (2022). Cyber risk and cybersecurity: a systematic review of data availability. Geneva Papers on Risk and Insurance - Issues and Practice. 47. 10.1057/s41288-022-00266-6.

Culot, G., Nassimbeni, G., Podrecca, M., & Sartor, M. (2021). The ISO/IEC 27001 information security management standard: Literature review and theory-based research agenda.

https://digital.lib.washington.edu/researchworks/bitstream/handle/1773/34826/Doyle%20-%20Capstone.pdf?sequence=1.

Fouad, N. S. (2021). Securing higher education against cyberthreats: from an institutional risk to a national policy challenge. *Journal of Cyber Policy*, *6*(2), 137–154. https://doi.org/10.1080/23738871.2021.1973526

Girgis, Sherry Gadallah, Mahmoud (2018) "Deep L earning Algorithms for Detecting Fake News in Online Text" Retrieved 2021-05-08

Ismail Fawaz, H., Forestier, G., Weber, J., Idoumghar, L., & Muller, P. A. (2019). Deep learning for time series classification: a review. Data Mining and Knowledge Discovery. https://doi.org/10.1007/s10618-019-00619-1

Itankan, W. A. (2023). The relationship between Teachers' Qualification and the impact on Student's Academic Achievements in Senior Secondary School Mathematics in Taraba state. International [1] Journal of Cognitive Research in Science, Social Sciences, & Law (IJCRSSSL), 3(2)

Lin, Yuanguo& Chen, Hong & Xia, Wei & Lin, Fan & Wu, Pengcheng& Wang, Zongyue& Li, Yong. (2023). A Comprehensive Survey on Deep Learning Techniques in Educational Data Mining. 10.48550/arXiv.2309.04761.

Máchová, R., &Lněnička, M. (2017). Evaluating the quality of open data portals on the national level. Journal of Theoretical and Applied Electronic Commerce Research. https://doi.org/10.4067/S0718-18762017000100003

Martey, P., Dominique, G. &Hassani, B. (2018). Learning Models, 1–19. https://doi.org/10.3390/risks6020038

Naagas, M. A., &Palaoag, T. D. (2018). A threat-driven approach to modeling a campus network security. In *Proceedings of the 6th International Conference on Communications and Broadband Networking*[1] (pp. 3193-3196). [2] [DOI: 10.1145/3193092.3193096]

Nainna, M. A., Bass, J., &Speakman, L. (2020). Cyber Threat Intelligence Sharing in Nigeria. California State University, San Bernardino

National Association of Corporate Directors. (2017). Cyber-risk oversight handbook (p. 16).

National Institute of Standards and Technology [NIST]. (n.d.). Cybersecurity Framework (CSF) 1.1. https://www.nist.gov/cyberframework

Provan, N. (2023). "Impact of School Funding on Student Achievement" (2023). *Capstone Projects and Master's Theses*. 1643.Retrieved from https://digitalcommons.csumb.edu/caps_thes_all/1643

Sean Brooks et al., (2017). An Introduction to Privacy Engineering and Risk Management in Federal Systems. http://nvlpubs.nist.gov/nistpubs/ir/2017/NIST.IR.8062.pdf.

Sidda, Sakunthala&Kiranmayi, R &Nagaraju, P. (2017). A Review on Artificial Intelligence Techniques in Electrical Drives Neural Networks, Fuzzy logic, and Genetic Algorithm. 10.1109/SmartTechCon.2017.8358335.

Tang, L., & Mahmoud, Q. H. (2021). A Novel Deep Learning Architecture for Phishing Website Detection.

The TQM Journal, 33(7), 76-105. https://doi.org/10.1108/TQM-09-2020-0202

The White House. (2024, May 7). Fact Sheet: 2024 Report on the Cybersecurity Posture of the United States [Website]. Retrieved from https://www.whitehouse.gov/oncd/briefing-room/2024/05/07/fact-sheet-cybersecurity-posture-report/

Wang, J., Neil, M., & Fenton, N.E. (2020). A Bayesian network approach for cybersecurity risk assessment implementing and extending the FAIR model. *Comput. Secur., 89*.

Weerakkody, V., Irani, Z., Kapoor, K., Sivarajah, U., &Dwivedi, Y. K. (2017). Open data and its usability: an empirical view from the Citizen's perspective. Information Systems Frontiers. https://doi.org/10.1007/s10796-016-9679-1

Weerakkody, V., Irani, Z., Kapoor, K., Sivarajah, U., &Dwivedi, Y. K. (2017). Open data and its usability: an empirical view from the Citizen's perspective. Information Systems Frontiers. https://doi.org/10.1007/s10796-016-9679-1

Zhang, C., & Zhang, L. (2021). Understanding the evolution of open government data research: Towards open data sustainability and smartness. International Review of Administrative Sciences,[1] 89(2), 2085232110099.