A Secure and Reliable Framework for Explainable Artificial Intelligence (XAI) in Smart City Applications

Mohammad Algarni
Department of Information Technology, College of Computer and Information Sciences, Majmaah University, Saudi Arabia
441104419@s.mu.edu.sa (corresponding author)

Shailendra Mishra
Department of Information Technology, College of Computer and Information Sciences, Majmaah University, Saudi Arabia
s.mishra@mu.edu.sa

Received: 29 April 2024 | Revised: 15 May 2024 | Accepted: 17 May 2024

 DOI: https://doi.org/10.48084/etasr.7676

ABSTRACT

Living in a smart city has many advantages, such as improved waste and water management, access to quality healthcare facilities, effective and safe transportation systems, and personal protection. Explainable AI (XAI) is called a system that is capable of providing explanations for its judgments or predictions. This term describes a model, its expected impacts, and any potential biases that may be present. XAI tools and frameworks can aid in comprehending and trusting the output and outcomes generated by machine-learning algorithms. This study used XAI methods to classify cities based on smart city metrics. The logistic regression method with LIME achieved perfect accuracy, precision, recall, and F1-score, predicting correctly all cases.

Keywords-machine learning; explainable artificial intelligence (XAI); smart city; artificial intelligence

I. INTRODUCTION

The emergence of smart cities marks a transformative era in urban development, leveraging innovative technologies to revolutionize city life. This evolution promises optimized waste and water management, superior healthcare, safer transportation, and increased personal security. Artificial Intelligence (AI) empowers data-driven decision-making processes to improve quality of life. However, understanding the rationale behind AI-driven decisions remains a fundamental challenge. As transparency is indispensable, understanding AI results is not optional but necessary [1]. This imperative has led to the development of Explainable Artificial Intelligence (XAI), to make AI models transparent and understandable to humans. XAI is vital for fostering trust and understanding in AI-driven solutions, particularly in legal and ethical contexts, as it goes beyond simple predictions and dives into reasoning. By elucidating the inner workings of AI models, XAI enables users to identify and address potential biases within the system, ensuring accountability and compliance [2]. Incorporating XAI into smart city applications presents challenges, such as security risks due to extensive data and communication networks, real-time data processing demands, and vulnerabilities in IoT devices, while ensuring the accountability of AI system operators [3].

This study investigates a secure and reliable framework for XAI in the context of smart city applications to promote the trustworthiness of AI systems and to ensure that their decisions are comprehensible and explainable to both city authorities and residents [4]. The primary goal is to ensure that AI systems are not only technologically sophisticated but also ethically responsible. Achieving this objective requires carefully balancing the use of AI transformative capabilities and the protection of the rights and well-being of individuals living in these urban settings. This study focuses on advancing technology and ensuring ethical considerations in the deployment of AI [5, 6] by:

- Investigating XAI systems for smart city environments to ensure transparency and comprehension of AI-driven decisions.
- Addressing concerns about accountability, transparency, and potential biases inherent in AI algorithms and decision-making processes.
- Fostering public trust and confidence in the reliability and efficacy of AI technologies used within smart city environments.

II. RELATED WORKS

Smart cities represent a paradigm shift in urban development, leveraging AI and advanced technologies to improve urban infrastructure and services. AI is widely used in various smart city applications, such as traffic management, healthcare, and energy optimization. However, the opacity of AI algorithms in these applications has led to a growing demand for transparency and accountability. Smart cities represent a transformative approach to urban development, harnessing the power of advanced technologies and data-driven solutions to optimize urban infrastructure and services. The integration of AI systems is central to this transformation and plays an essential role in enhancing various aspects of city life. These AI-driven solutions are applied in various domains such as traffic management, healthcare energy management, public safety, etc. [7-19].

XAI traffic management systems involve traffic flow optimization in smart cities. The importance of transparent algorithms to provide real-time explanations for traffic decisions has been highlighted, making it easier for city planners and residents to understand and trust the system [13]. Public transport plays a vital role in smart city traffic management and transparent AI models are essential in this context to gain public acceptance and cooperation [16, 18, 19]. XAI systems in healthcare involve patient diagnostics and play a significant role in providing interpretable and transparent diagnostic results. In addition, the AI and XAI models must protect patient privacy. Integration of AI and data-driven technologies into smart cities has opened numerous opportunities to improve urban living. However, this transformation has also exposed smart city applications to a range of security challenges, as AI becomes more deeply embedded in the fabric of urban infrastructure and services. Ensuring the confidentiality, integrity, and availability of data and systems becomes paramount. For instance, in [16], security concerns in smart cities were emphasized, including data privacy and potential cyber threats to critical infrastructure, calling for secure AI systems to protect sensitive information.

Several research gaps have emerged in the context of XAI and its integration into smart city applications. First, it is necessary to quantitatively investigate the impact of security measures on XAI, providing a deeper understanding of the tradeoffs between security and the effectiveness of XAI. Second, while research has provided a wide overview of XAI in smart cities, there is a distinct lack of domain-specific research, necessitating focused investigations on healthcare, traffic management, and energy optimization. In addition, the development of real-time XAI solutions, particularly for time-sensitive applications, represents a pressing gap in the current research landscape. Furthermore, the role of emerging technologies, such as quantum computing and advanced cryptography, and their influence on AI/XAI and security within smart cities remains largely unexplored.

III. METHOD

This study used a smart city index dataset [20]. The initial steps involved data cleaning to remove any inconsistencies or missing values. Then, feature extraction was performed, focusing on key aspects of smart cities, including Smart Mobility (e.g., traffic flow and public transportation efficiency), Smart Environment (e.g., air quality and waste management), Smart Government (e.g., e-governance services), Smart Economy (e.g., economic growth indicators), Smart People (e.g., education and citizen engagement), and Smart Living (e.g., healthcare and quality of life). Each of these features was assigned a weight based on its importance in contributing to overall smart city performance. These weighted features were then aggregated using a specified formula to calculate the Smart City Index. This index provides a composite score that reflects each city's overall performance and progress in becoming smarter and more technologically advanced. The resulting Smart City Index allows for the classification of cities into different performance levels, facilitating comparative analysis and targeted improvements.

Two XAI algorithms were employed: Random Forest (RF) with Shapley Additive Explanations (SHAP) and Linear Regression (LR) with Local Interpretable Model-agnostic Explanations (LIME). The application of SHAP and LIME facilitates the interpretation of model decisions, shedding light on the factors influencing results and improving the transparency of the smart city framework. Figure 1 shows a detailed analysis of the method.

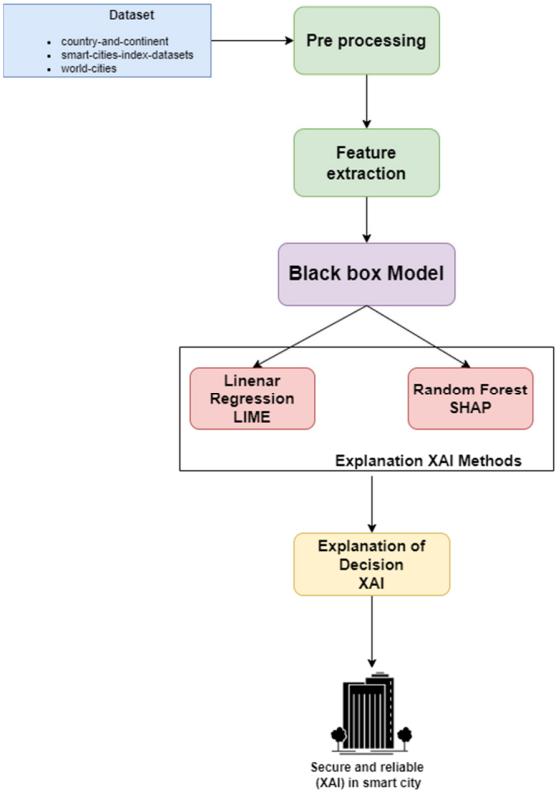

Fig. 1. Research framework.

IV. IMPLEMENTATION

This study used the Python programming language and Jupyter Notebook to seamlessly integrate ML models with

cutting-edge XAI techniques, such as SHAP and LIME. The key components of the system and their interconnections are described below.

A. Core Components

- Smart City Data Platform: The smart city dataset serves as the foundation for data-driven decision-making [20].
- ML Models: The ML modules are responsible for developing and deploying predictive models trained on historical and real-time data.
- XAI Integration Layer: This layer acts as a bridge between ML models and interpretability techniques. It facilitates the seamless incorporation of XAI methods into the decision-making process, ensuring that model outputs are understandable and justifiable.

B. Integration of XAI Techniques

1) SHAP

SHAP values are computed for each feature in the ML model, quantifying its contribution to the model output. These values provide a nuanced understanding of feature importance, allowing stakeholders to discern which features significantly influence the model's predictions. By integrating SHAP values into the XAI layer, the framework offers a comprehensive explanation of the underlying decision-making process.

2) LIME

LIME generates local approximations of the decision boundaries of ML models. These approximations are tailored to interpret individual predictions, offering insights into the factors that influence specific outcomes. By providing interpretable explanations at the instance level, LIME facilitates a deeper understanding of how the model arrives at its predictions, thus fostering trust and confidence in AI-driven decision-making.

C. Dataset

Data from [17] were used, which are based on the Smart City Index, a well-established framework that is used to analyze and benchmark the effectiveness of smart city initiatives. Cities with exemplary performance in smart city dimensions were classified as high performance. These cities exhibit robust infrastructure, efficient services, and innovative initiatives that contribute to their overall advancement. Cities with moderate scores, indicating satisfactory but not exceptional performance, were classified as medium performance. These cities may have areas of improvement but generally demonstrate competence in smart city development. On the other hand, cities with lower scores, indicating significant room for improvement in smart city initiatives, were categorized as low performance. The dataset was split into training, validation, and test sets in a 70:15:15 ratio.

D. Data Visualizations

Figure 2 shows Smart City Index totals, which is a detailed metric on the efficacy of smart city efforts in various metropolitan areas. This index provides a comprehensive perspective of a city's progress toward becoming smarter and more technologically sophisticated, considering factors such as Smart Mobility, Smart Environment, Smart Government, Smart Economy, Smart People, and Smart Living. Cities can assess themselves against global norms using this index, creating healthy competition and driving continuous infrastructure improvements. Stakeholders, legislators, and people may use this index to get useful insights into each city's development strengths and areas for improvement. Figure 3 shows the distribution of smart city components, such as Smart Mobility, Smart Environment, Smart Government, etc., in different cities or countries. These visualizations help stakeholders swiftly grasp essential information and identify trends, resulting in a more comprehensive knowledge of the smart city scene.

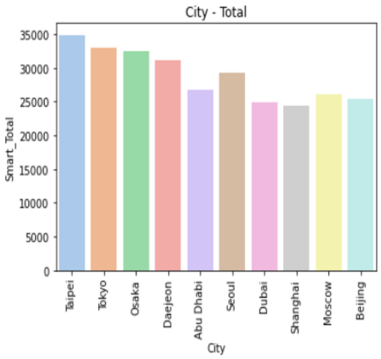

Fig. 2. Total smart city index.

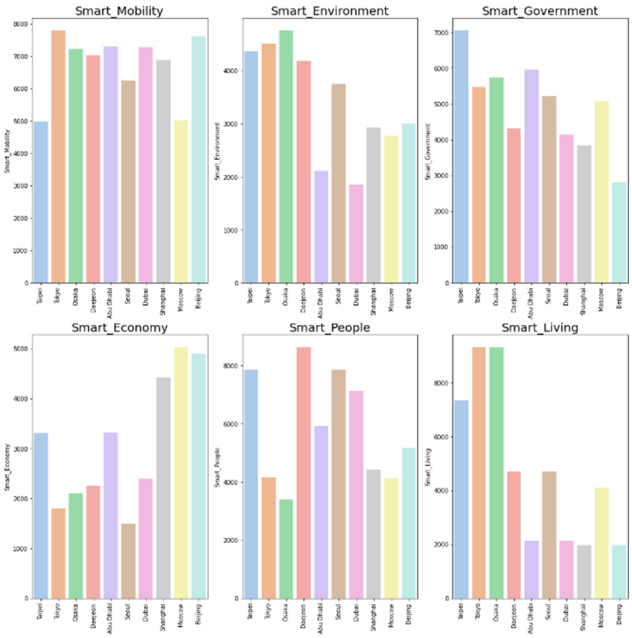

Fig. 3. Smart index bar plot per city.

Figure 4 shows a heatmap that visualizes distinct smart city characteristics, offering a fast understanding of the dataset's linkages and dependencies. The color intensity reflects the degree and direction of correlations, assisting in the

identification of patterns and insights across many smart city characteristics.

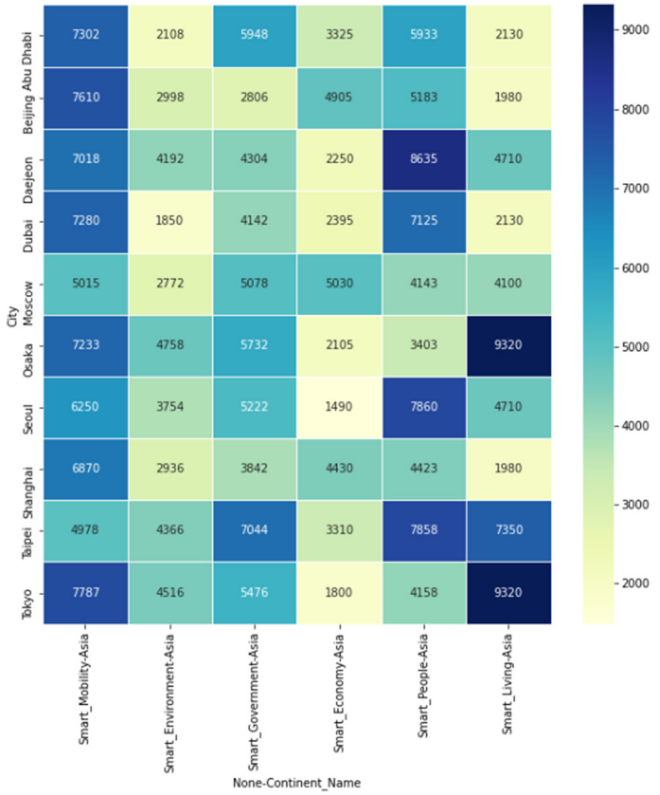

Fig. 4. Heatmap for cities and smart city characteristics.

E. ML Implementation

The ML models were trained using smart city metrics data, including features such as Smart Mobility, Smart Environment, Smart_Government, Smart Economy, Smart People, and Smart Living. The models performed classification to predict city performance across these metrics, categorizing cities into three performance levels, high, medium, and low, based on their aggregated scores. Detailed preprocessing ensured consistency and reliability in model training and evaluation.

1) Logistic Regression (LR) Training Process for LIME

Given the simplicity of LR, standard hyperparameter settings were used, focusing on regularization strength and solver choice. The model was trained on labeled data, and its results were used for LIME explanations. Figures 5 and 6 show that LIME indicates a 0.69 probability for Smart_Living to influence the total Smart_Index. In addition, feature contributions show the positive impact of Smart Living, Smart Environment, Smart People, and Smart Mobility, and the negative impact of Smart Government and Smart Economy. The prediction probabilities, highlighted conditions, and feature values offer transparency into the decision-making process of the model, in predicting the Smart_Total class. This level of interpretability is crucial to building trust and understanding. However, the evaluation of the model's overall effectiveness should consider global performance metrics, alignment with application objectives, and user feedback. To improve transparency and reliability, ongoing refinement and user interaction are essential to ensure the practical utility and ethical implementation of the model. Figures 5 and 6 show a summary of the influential features and their contributions to the model. Positive numbers indicate a positive influence on the projected outcome, whereas negative values suggest a negative impact.

Logistic Regression provided a baseline for classifying smart city performance metrics, yielding interpretable coefficients that indicate the influence of each feature. However, LIME was employed to enhance local interpretability and understand individual predictions. LIME allowed us to break down predictions for specific instances, offering a clearer, instance-level explanation of why a particular city was classified in a certain performance tier, thus improving the transparency and trust in model decisions.

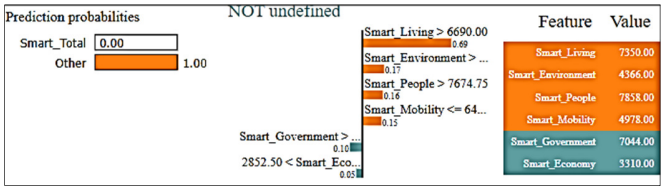

Fig. 5. Output of LR with LIME.

```
[43]: [('Smart_Living > 6690.00', 0.695231614111048),
       ('Smart_Environment > 4322.50', 0.19773913192039785),
       ('Smart_People > 7674.75', 0.1857830493147934),
       ('Smart_Mobility <= 6405.00', 0.13647081444088088),
       ('Smart_Government > 5668.00', -0.09185239782868802),
       ('2852.50 < Smart_Economy <= 4153.75', -0.039133080500545946)]
```

Fig. 6. LIME results.

2) Random Forest (RF) Regression with SHAP

The Smart Mobility, Smart Environment, Smart_Government, Smart Economy, Smart People, and Smart Living features were also selected for RF-SHAP. The number of trees, depth of trees, and feature split criteria parameters were optimized using grid search. The RF model was trained on the selected features, capturing complex relationships within the data. Figures 7 and 8 display the output of the RF algorithm with the SHAP values for the first ten occurrences in the test set, revealing how each feature contributes to the model results.

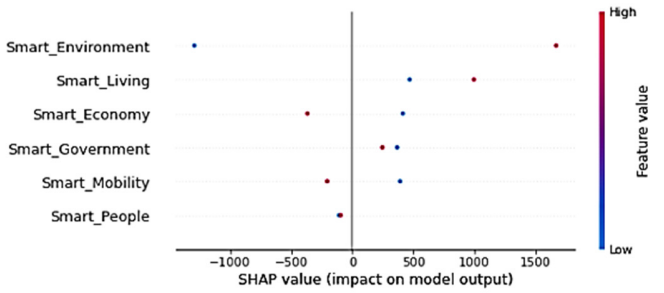

Fig. 7. Output for RF SHAP.

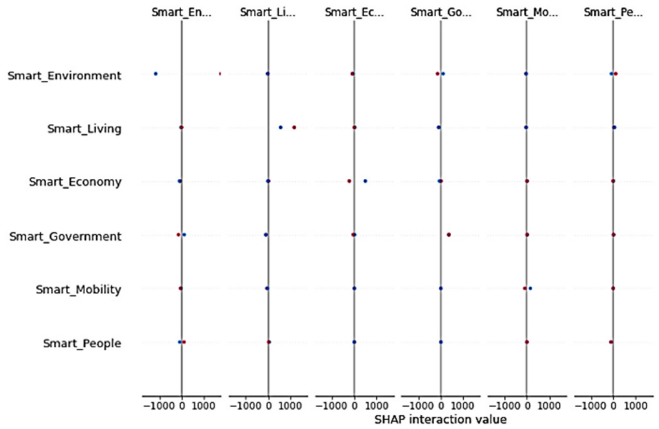

Fig. 8. SHAP interaction values.

V. RESULTS AND DISCUSSION

A. Results

The two XAI ML models were used to classify cities into three classes (high, medium, and low) according to their smart city features. Classification results were evaluated using key performance metrics, such as accuracy, precision, recall, and F1-score. The results shown in Table I indicate that the LR model performed flawlessly on the test set, as its accuracy, precision, recall, and F1-score were 99.9%. This means that the model correctly predicted all cases, with no false positives, false negatives, or misclassifications.

TABLE I. MODEL PERFORMANCE

Model	Accuracy	Precision	Recall	F1-score
SHAP	85%	82.3%	85.2%	82.3%
LIME	99.9%	99.9%	99.9%	99.9%

SHAP values help uncover the contribution of each feature to model predictions, offering a nuanced understanding of feature importance. On the other hand, LIME provides locally trustful explanations for individual predictions, enhancing the transparency of the model's decision-making process.

B. Discussion

Using XAI, stakeholders gain a transparent and interpretable view of the factors that influence the smart city classification. This clarity empowers decision-makers to navigate the complexities of smart city scenarios with enhanced confidence and reliability. These results serve as a comprehensive exploration of the multifaceted outcomes arising from the implemented XAI frameworks, offering valuable insights that transcend traditional model evaluation. Integrating interpretability tools ensures a holistic understanding of smart city models, fostering a more informed decision-making paradigm. XAI refers to a set of techniques and methodologies that aim to enhance the transparency and interpretability of ML models, particularly those used in complex systems. Traditional ML models, such as deep neural networks, often operate as black boxes, making it challenging to understand how they arrive at specific decisions. This lack of transparency can be a significant barrier, especially in critical applications such as smart cities, where decisions can have far-reaching implications for citizens and infrastructure. This study adopted a versatile approach employing two distinct learning models, namely LR and RF. This combination allows capturing different aspects of the underlying data patterns in smart cities to enhance interpretability and transparency. Two state-of-the-art explainability methods were used with LR and RF, LIME and SHAP, respectively. LIME facilitates local interpretability by providing insights into individual predictions, while SHAP offers a broader understanding of feature contributions across the entire dataset. This dual model and dual explanation strategy aimed to provide a comprehensive and nuanced view of the smart city framework, contributing to both accuracy and interpretability in decision-making processes.

VI. CONCLUSIONS

This study used LR and RF ML models in conjunction with LIME and SHAP, respectively, to improve decision-making processes in complex urban settings, promoting transparency and understanding. However, the study acknowledges several limitations, including constraints related to data availability, model specificity, interpretability challenges, and the dynamic nature of smart city ecosystems. Future work should explore advanced ML models and XAI techniques, integrate real-time data streams, address cybersecurity concerns, and extend the framework to other smart city domains, such as traffic management, healthcare, and energy optimization. Longitudinal studies, ethical considerations, and global collaboration efforts will be crucial to ensuring the responsible deployment and continuous improvement of XAI in smart city environments, fostering innovation and sustainable urban development.

REFERENCES

[1] V. Hassija *et al.*, "Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence," *Cognitive Computation*, vol. 16, no. 1, pp. 45–74, Jan. 2024, https://doi.org/10.1007/s12559-023-10179-8.

[2] I. D. Apostolopoulos and P. P. Groumpos, "Fuzzy Cognitive Maps: Their Role in Explainable Artificial Intelligence," *Applied Sciences*, vol. 13, no. 6, Jan. 2023, Art. no. 3412, https://doi.org/10.3390/app13063412.

[3] Z. Ullah, F. Al-Turjman, L. Mostarda, and R. Gagliardi, "Applications of Artificial Intelligence and Machine learning in smart cities," *Computer Communications*, vol. 154, pp. 313–323, Mar. 2020, https://doi.org/10.1016/j.comcom.2020.02.069.

[4] M. Schnieder, "Using Explainable Artificial Intelligence (XAI) to Predict the Influence of Weather on the Thermal Soaring Capabilities of Sailplanes for Smart City Applications," *Smart Cities*, vol. 7, no. 1, pp. 163–178, Feb. 2024, https://doi.org/10.3390/smartcities7010007.

[5] C. I. Nwakanma *et al.*, "Explainable Artificial Intelligence (XAI) for Intrusion Detection and Mitigation in Intelligent Connected Vehicles: A Review," *Applied Sciences*, vol. 13, no. 3, Jan. 2023, Art. no. 1252, https://doi.org/10.3390/app13031252.

[6] M. Ahmed, S. R. Islam, A. Anwar, N. Moustafa, and A. S. K. Pathan, Eds., *Explainable Artificial Intelligence for Cyber Security: Next Generation Artificial Intelligence*. Springer International Publishing, 2022.

[7] S. K. Jagatheesaperumal, Q.-V. Pham, R. Ruby, Z. Yang, C. Xu, and Z. Zhang, "Explainable AI Over the Internet of Things (IoT): Overview, State-of-the-Art and Future Directions," *IEEE Open Journal of the Communications Society*, vol. 3, pp. 2106–2136, 2022, https://doi.org/10.1109/OJCOMS.2022.3215676.

[8] C. Hurter *et al.*, "Usage of more transparent and explainable conflict resolution algorithm: air traffic controller feedback," *Transportation*

Research Procedia, vol. 66, pp. 270–278, Jan. 2022, https://doi.org/10.1016/j.trpro.2022.12.027.

[9] Z. A. E. Houda, B. Brik, and L. Khoukhi, "'Why Should I Trust Your IDS?': An Explainable Deep Learning Framework for Intrusion Detection Systems in Internet of Things Networks," *IEEE Open Journal of the Communications Society*, vol. 3, pp. 1164–1176, 2022, https://doi.org/10.1109/OJCOMS.2022.3188750.

[10] O. Loyola-González, "Understanding the Criminal Behavior in Mexico City through an Explainable Artificial Intelligence Model," in *Advances in Soft Computing*, Xalapa, Mexico, 2019, pp. 136–149, https://doi.org/10.1007/978-3-030-33749-0_12.

[11] K. A. Eldrandaly, M. Abdel-Basset, M. Ibrahim, and N. M. Abdel-Aziz, "Explainable and secure artificial intelligence: taxonomy, cases of study, learned lessons, challenges and future directions," *Enterprise Information Systems*, Sep. 2023, https://doi.org/10.1080/17517575.2022.2098537.

[12] P. Weber, K. V. Carl, and O. Hinz, "Applications of Explainable Artificial Intelligence in Finance—a systematic review of Finance, Information Systems, and Computer Science literature," *Management Review Quarterly*, vol. 74, no. 2, pp. 867–907, Jun. 2024, https://doi.org/10.1007/s11301-023-00320-0.

[13] M. M. Karim, Y. Li, and R. Qin, "Toward Explainable Artificial Intelligence for Early Anticipation of Traffic Accidents," *Transportation Research Record*, vol. 2676, no. 6, pp. 743–755, Jun. 2022, https://doi.org/10.1177/03611981221076121.

[14] Z. Li, Y. Zhu, and M. Van Leeuwen, "A Survey on Explainable Anomaly Detection," *ACM Transactions on Knowledge Discovery from Data*, vol. 18, no. 1, Jun. 2023, Art. no. 23, https://doi.org/10.1145/3609333.

[15] A. Rawal, J. McCoy, D. B. Rawat, B. M. Sadler, and R. St. Amant, "Recent Advances in Trustworthy Explainable Artificial Intelligence: Status, Challenges, and Perspectives," *IEEE Transactions on Artificial Intelligence*, vol. 3, no. 6, pp. 852–866, Dec. 2022, https://doi.org/10.1109/TAI.2021.3133846.

[16] A. Procopiou and T. M. Chen, "Explainable AI in Machine/Deep Learning for Intrusion Detection in Intelligent Transportation Systems for Smart Cities," in *Explainable Artificial Intelligence for Smart Cities*, CRC Press, 2021.

[17] D. Prabakar, M. Sundarrajan, S. Prasath Alias Surendhar, M. Ramachandran, and D. Gupta, "Trust Model Based Data Fusion in Explainable Artificial Intelligence for Edge Computing Using Secure Sequential Discriminant Auto Encoder with Lightweight Optimization Algorithm," in *Explainable Edge AI: A Futuristic Computing Perspective*, A. E. Hassanien, D. Gupta, A. K. Singh, and A. Garg, Eds. Springer International Publishing, 2023, pp. 139–160.

[18] I. Batra, A. Malik, S. Sharma, C. Sharma, and S. Hosen, "Explainable Artificial Intelligence into Cyber-Physical System Architecture of Smart Cities: Technologies, Challenges, and Opportunities," *Journal of Electrical Systems*, vol. 20, no. 2, pp. 2343–2362, Apr. 2024, https://doi.org/10.52783/jes.2000.

[19] M. H. Kabir, K. F. Hasan, M. K. Hasan, and K. Ansari, "Explainable Artificial Intelligence for Smart City Application: A Secure and Trusted Platform," in *Explainable Artificial Intelligence for Cyber Security: Next Generation Artificial Intelligence*, M. Ahmed, S. R. Islam, A. Anwar, N. Moustafa, and A.-S. K. Pathan, Eds. Springer International Publishing, 2022, pp. 241–263.

[20] M. Monteiro, "Smart Cities Index Datasets." Kaggle, [Online]. Available: https://www.kaggle.com/datasets/magdamonteiro/smart-cities-index-datasets.