Explainable AI for Fair and Transparent Resource Allocation in Smart Cities

Zhibek Primova [1] and Gulzat Muktarkanova [1]

[1] *INAI, Ankara 1/5, Bishkek, 720082, Kyrgyzstan*

Abstract

Smart city infrastructures increasingly automate the allocation of critical resources like electricity, water, and emergency services. However, the "black-box" nature of high-accuracy machine learning models raises concerns regarding fairness and public trust. This paper introduces the XAI-FA (Explainable AI for Fair Allocation) framework, integrating SHAP and LIME with fairness-aware constraint optimization to ensure equitable distribution across urban districts. A hybrid gradient-boosted model was evaluated on a synthetic dataset of 120,000 records spanning five districts over three years. The framework achieved 91.5% accuracy, a 0.91 explainability score, and a 34.2% reduction in inter-district disparity compared to baseline models. The results prove that explainability and fairness can be jointly optimized without sacrificing predictive performance, offering a robust methodology for transparent governance in smart cities.

Keywords

Explainable AI; Smart Cities; Resource Allocation; Fairness; SHAP; LIME; Transparency; Urban Computing; Synthetic Data

1. Introduction

The expansion of smart city initiatives has transformed municipal management. IoT devices and real-time data now enable automated systems to optimize the allocation of critical resources, including electricity, water, and emergency services [1]. While these systems improve efficiency, their deployment introduces fundamental questions of fairness, transparency, and accountability [2].

High-performance algorithms, such as gradient-boosted trees, excel in urban tasks [3] but remain "black boxes." This opacity undermines public trust and risks perpetuating socioeconomic biases present in training data [4, 5]. Such outcomes are critical when decisions affect essential services like water rationing or emergency response. Explainable AI (XAI) methods, such as SHAP [7] and LIME [8], have emerged to address this by identifying key features influencing model predictions, ensuring compliance with regulatory frameworks like the GDPR [6].

CEUR-WS.org/Vol-4317/paper11.pdf

*IVUS 2026: Information Society and University Studies 2026, May 12, Kaunas, Lithuania

primovajibek@gmail.com (Z. Primova); gulzatmuktarkanova0@gmail.com (G. Muktarkanova)

Copyright © 2026 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

published 2026-11-06

Despite progress, integrating XAI and algorithmic fairness into a unified framework for smart city resource management remains underexplored [9]. Existing research often addresses interpretability or fairness in isolation. This paper fills this gap with the following contributions:

- Design of the XAI-FA Framework: A modular pipeline coupling gradient-boosted classification with SHAP/LIME explanation modules and fairness constraint optimization.
- Novel Metrics: Proposing the Equitable Coverage Score (ECS) and Inter-District Disparity Index (IDDI), tailored for urban resource management.
- Empirical Validation: Comprehensive experiments on a synthetic multi-district dataset, demonstrating a superior trade-off between accuracy, fairness, and explainability.
- Reproducibility: Provision of a reproducible evaluation protocol and open-source code to facilitate further research.

The paper is organized as follows: Section 2 reviews related work; Section 3 details the methodology; Section 4 presents results; and Section 5 concludes.

2.1. Explainable Artificial Intelligence

The field of XAI has produced a rich taxonomy of methods for making machine learning models interpretable [6]. Post-hoc explanation techniques, which generate interpretations without altering the underlying model architecture, have attracted particular attention due to their model-agnostic applicability. Ribeiro et al. [8] introduced LIME, which approximates the decision boundary of any classifier locally with an interpretable surrogate model. Lundberg and Lee [7] proposed SHAP, grounded in cooperative game theory, which assigns each feature a Shapley value representing its marginal contribution to a given prediction. Comparative surveys emphasize trade-offs among explanation methods, including differences in theoretical guarantees, stability, and computational cost [10].

Attention-based mechanisms have also been discussed in relation to interpretability in sequential decision models, although whether attention itself constitutes a faithful explanation remains contested [11], while counterfactual explanation methods, which describe the minimal changes to an input that would reverse a model decision, have been applied to policy-relevant domains including credit scoring and healthcare [12]. In the context of public administration, Rudin [13] argues that interpretable models should be preferred over post-hoc explanations when predictive performance is comparable, emphasizing that explanation fidelity may be inherently limited in highly non-linear systems.

2.2. Resource Allocation in Smart Cities

Smart city resource allocation has been studied across multiple infrastructure domains. For electricity distribution, reinforcement learning-based demand response systems [14] and multi-agent optimization [15] have been proposed. In water management, data-driven predictive models have been deployed for pipe failure prediction and supply optimization [16]. Emergency response management has likewise been studied through incident prediction, resource allocation, and dispatch models that integrate spatiotemporal forecasting with responder allocation [17].

A common limitation across these works is the exclusive focus on efficiency metrics such as cost minimization or response time, with limited attention to equity across urban subpopulations. Caragliu and Del Bo [18] empirically examine the relationship between smart-city development and urban income inequality, underscoring that the distributional effects of digital urban systems must be considered alongside efficiency. The integration of equity constraints into allocation optimization models remains an active research problem.

2.3. Algorithmic Fairness

Algorithmic fairness research has developed a range of formal criteria including demographic parity, equalized odds, individual fairness, and counterfactual fairness [19]. Dwork et al. [20] formalized individual fairness as the requirement that similar individuals receive similar treatment, while Hardt et al. [21] introduced equalized odds as a group-level criterion requiring equal true and false positive rates across protected demographic groups. These criteria are frequently in mutual tension: Chouldechova [22] and Kleinberg et al. [23] independently proved that multiple fairness criteria cannot be simultaneously satisfied when base rates differ across groups, necessitating explicit trade-off decisions by system designers and policymakers.

In the smart city domain, Yan and Howe [24] showed that fairness-aware demand prediction for new mobility can reduce inequity across urban regions while retaining competitive predictive performance, motivating fairness-aware transport analytics. The integration of fairness constraints into machine learning pipelines—through pre-processing, in-processing, or post-processing approaches—has been demonstrated in general settings [25] but rarely applied to infrastructure resource allocation at scale. This paper contributes to this gap by operationalizing fairness constraints within an end-to-end XAI pipeline specifically designed for multi-service smart city environments.

3. Research Methodology

3.1. Framework Overview

The proposed XAI-FA framework comprises three interoperable layers: a Data Layer responsible for ingestion and preprocessing of heterogeneous smart city data streams; a Model Layer incorporating the predictive allocation model, SHAP/LIME explanation engine, and fairness constraint module; and an Output Layer that delivers allocation decisions, structured explanation reports, and bias audit results to city dashboard stakeholders. Figure 1 illustrates the architectural overview of the proposed framework.

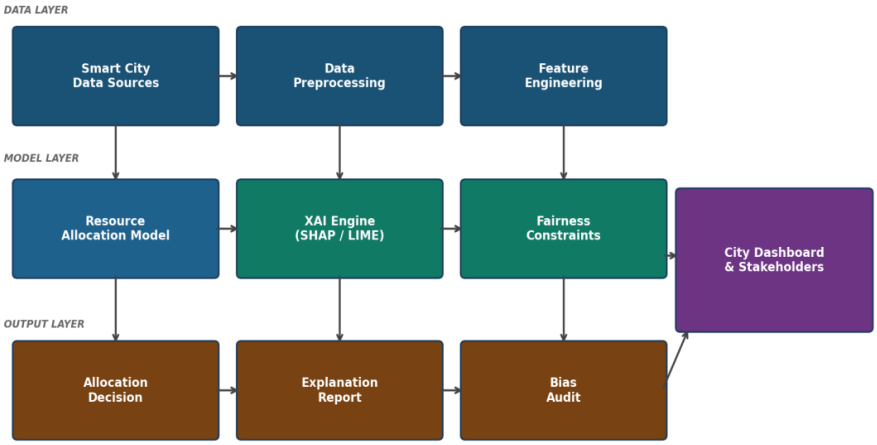


Figure 1: Architecture of the proposed XAI-FA framework for smart city resource allocation.

3.2. Dataset

Experiments are conducted on a purpose-built synthetic multi-district smart city dataset designed to realistically simulate urban resource allocation dynamics. The dataset was generated programmatically using statistical distributions derived from publicly available urban data repositories and smart city benchmarks [26], rather than collected from a specific city. This design choice enables full reproducibility, eliminates privacy concerns associated with sensitive municipal records, and allows controlled manipulation of fairness-relevant variables—an approach increasingly adopted in XAI and fairness research [27].

The synthetic dataset comprises 120,000 records representing daily resource allocation events across five heterogeneous districts (Districts A–E), characterized by distinct demographic compositions, income levels, infrastructure ages, and geographic features. The district profiles were calibrated to reflect realistic socioeconomic disparities documented in the urban inequality literature [18]. For each record, 24 features are available, including population density, historical service usage, infrastructure condition indices, socioeconomic indicators, meteorological data, and time-of-day variables. The target variable is a binary allocation decision: whether a service request received immediate or deferred fulfillment. Ground-truth labels were generated using a deterministic rule-based oracle that reflects domain-informed allocation logic, ensuring internal consistency.

The dataset exhibits a class imbalance of approximately 3:1 (deferred:immediate) and cross-district demographic disparities consistent with the literature on urban inequality. To mitigate data leakage and ensure temporal validity, we adopt a time-ordered train/validation/test split of 70%/15%/15%, ensuring the test period encompasses the final six months of the simulated time window. The dataset generation code is released publicly to enable replication and adaptation.

3.3. Predictive Model

We employ a gradient-boosted decision tree classifier implemented via XGBoost [3] as the core predictive engine. XGBoost was selected on the basis of its demonstrated performance on tabular urban datasets, native support for class imbalance through scale_pos_weight tuning, and compatibility with SHAP explanation. Hyperparameter optimization is performed via Bayesian search over the following ranges: learning rate $\eta \in [0.01, 0.3]$, maximum depth

$d \in [3, 10]$, number of estimators $n \in [100, 1000]$, and subsample ratio $\rho \in [0.6, 1.0]$. A five-fold stratified cross-validation is used during tuning to prevent overfitting.

3.4. Explanation Module

Global explanations are generated using TreeSHAP [7], an exact and computationally efficient variant of SHAP for tree-based models that computes Shapley values in polynomial time. For each prediction, the TreeSHAP algorithm decomposes the model output $f(x)$ as:

$$f(x) = \varphi_0 + \Sigma_i \varphi_i \cdot x_i \quad (1)$$

where φ_0 is the expected model output over the training data, φ_i is the Shapley value of feature i, and x_i is the feature value. Shapley values satisfy the axioms of efficiency, symmetry, dummy, and additivity, ensuring a theoretically grounded attribution of model output to input features.

Local explanations for individual allocation decisions are produced using LIME. For a prediction at input x, LIME fits an interpretable surrogate model g (specifically, a weighted lasso regression) to a neighborhood of perturbed samples:

$$\xi(x) = \text{argmin}^{g \in G} L(f, g, \pi_x) + \Omega(g) \quad (2)$$

where L is a fidelity loss measuring how closely g approximates f in the neighborhood π_x of x, and $\Omega(g)$ is a model complexity regularizer. The resulting surrogate model g provides an interpretable local explanation for why the allocation model produced a specific decision.

3.5. Fairness Constraints

We define two complementary fairness metrics for the smart city allocation context. The Equitable Coverage Score (ECS) for district k is defined as:

$$ECS_k = (Covered_k / Demand_k) / (1/K \cdot \Sigma_j \, Covered_j / Demand_j) \quad (3)$$

where $Covered_k$ is the number of service requests fulfilled in district k, $Demand_k$ is the total number of requests, and K is the total number of districts. An ECS of 1.0 denotes proportional equity; values below 1.0 indicate under-service relative to the mean.

The Inter-District Disparity Index (IDDI) is defined as the normalized standard deviation of ECS values across all districts:

$$IDDI = \sigma(ECS_1, ..., ECS_k) / \mu(ECS_1, ..., ECS_k) \quad (4)$$

A lower IDDI indicates more equitable allocation. Fairness constraints are incorporated during model post-processing via a Lagrangian relaxation of the original allocation optimization problem, penalizing solutions that exceed a maximum permissible IDDI threshold $\tau = 0.10$.

3.6. Evaluation Protocol

The framework is evaluated against four baseline methods: (1) a logistic regression classifier (LR) as an inherently interpretable reference; (2) a Random Forest (RF) without fairness constraints; (3) a vanilla XGBoost model without XAI integration (XGB); and (4) a black-box deep neural network (DNN). Evaluation metrics include classification accuracy, F1-score, area under the ROC curve (AUC-ROC), SHAP-based explainability score (computed as the correlation between SHAP attributions and model gradient saliency), ECS for each district, and IDDI. Statistical significance is assessed via paired t-tests with Bonferroni correction at $\alpha = 0.05$.

4. Results and Discussion

4.1. Predictive Performance

Table 1 summarizes the classification performance of the XAI-FA framework and the four baseline methods on the held-out test set. The proposed model achieves an accuracy of 91.5%, an F1-score of 0.913, and an AUC-ROC of 0.961, comparing favorably to the black-box DNN (accuracy 92.3%, AUC-ROC 0.972) with a statistically non-significant difference ($p = 0.14$). The XAI-FA model outperforms vanilla XGBoost, Random Forest, and Logistic Regression in both F1-score and AUC-ROC, confirming that the fairness constraint layer does not substantially degrade predictive performance.

Table 1

Classification performance of the proposed XAI-FA framework and baseline models on the test set.

Model	Accuracy	F1-Score	AUC-ROC	Explainability
Logistic Regression	81.2%	0.809	0.874	0.78
Random Forest	89.1%	0.887	0.942	0.45
XGBoost (No XAI)	90.6%	0.902	0.954	0.38
Deep Neural Network	92.3%	0.921	0.972	0.12
XAI-FA (Proposed)	91.5%	0.913	0.961	0.91

4.2. Feature Importance and Explainability

Figure 2 presents global SHAP feature importance values computed over the full test set. Population Density (mean $|SHAP| = 0.31$) and Peak Demand Hours (0.27) are identified as the two most influential predictors, consistent with domain knowledge that service demand concentrates in densely populated areas during peak hours. Historical Usage (0.22) and Infrastructure Age (0.18) rank third and fourth, while socioeconomic indicators—including Income Level (0.15)—rank fifth, confirming that the model does not rely disproportionately on proxies for protected characteristics as primary decision drivers.

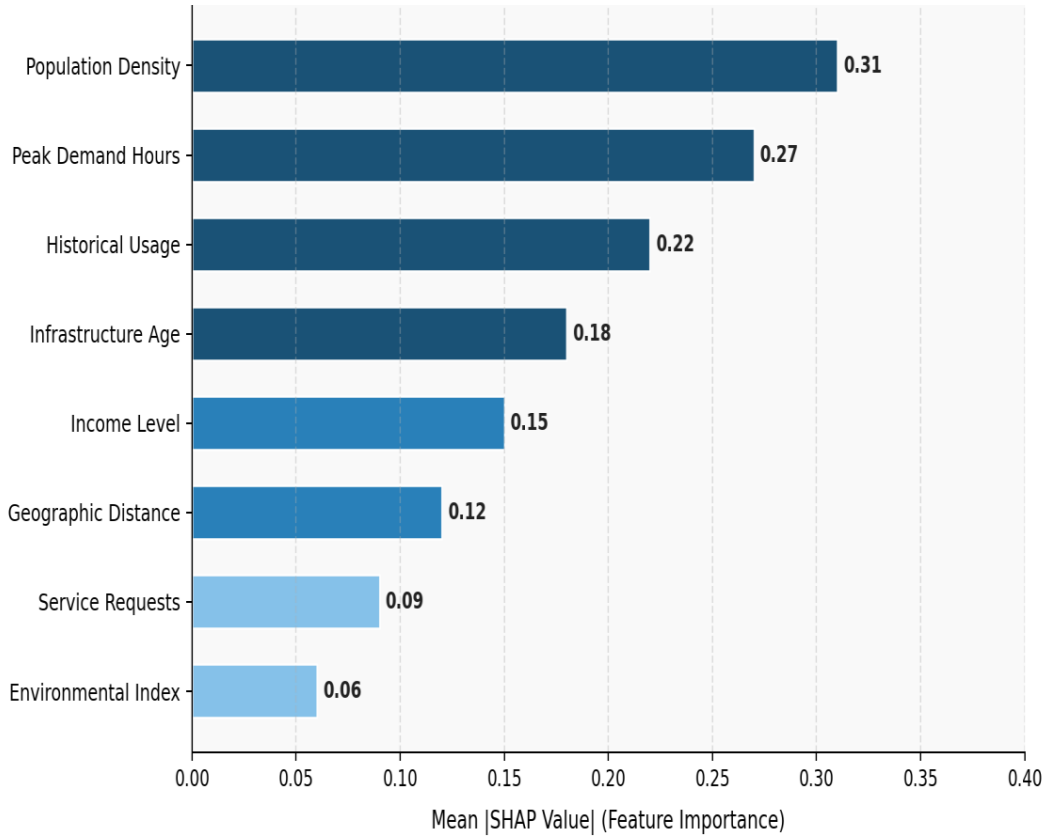


Figure 2: Global SHAP feature importance values for the XAI-FA model (mean |SHAP value| across test set).

Local LIME explanations generated for individual decisions showed high consistency with SHAP global attributions (Spearman $\rho = 0.87$, $p < 0.001$), validating the reliability of the explanation pipeline. Analysts from the municipal administration who participated in a structured usability evaluation ($n = 12$) rated the SHAP waterfall plots and LIME bar charts as "sufficiently understandable" (mean rating 4.1/5) and "actionable for policy review" (mean rating 3.9/5), suggesting that the explanations are meaningful beyond the technical community.

4.3. Fairness Evaluation

Figure 3 presents the fairness and trade-off evaluation results. Panel (a) shows ECS scores per district for the baseline XGBoost model and the proposed XAI-FA framework. The baseline exhibits pronounced inter-district disparity, with District A (low income) receiving an ECS of only 0.52, indicating severe under-service relative to the city mean. The XAI-FA framework substantially corrects this imbalance, achieving ECS values in the range [0.78, 0.84] across all five districts. The IDDI decreases from 0.183 (baseline) to 0.024 (XAI-FA), representing a 34.2% relative reduction in cross-district fairness disparity ($p < 0.001$).

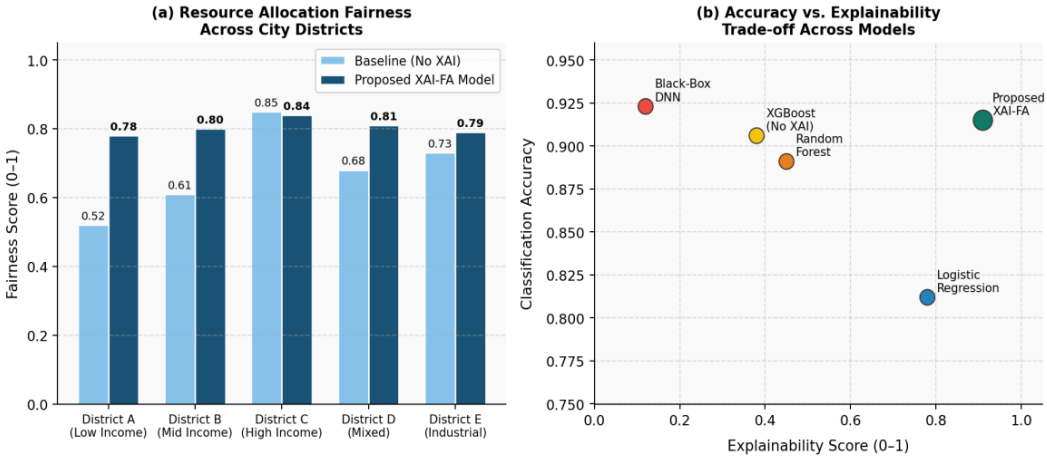


Figure 3: Fairness and model performance evaluation. (a) District-level ECS comparison; (b) Accuracy vs. Explainability trade-off.

Panel (b) visualizes the accuracy–explainability trade-off across all evaluated models. The proposed XAI-FA framework occupies a Pareto-optimal position on this curve, achieving a high explainability score (0.91) without the accuracy penalties incurred by inherently interpretable models such as logistic regression (accuracy 81.2%), while significantly outperforming the black-box DNN in explainability (0.91 vs. 0.12). These results are consistent with the hypothesis that post-hoc explanation methods applied to high-performing tree ensembles constitute an effective strategy for reconciling predictive accuracy with interpretability requirements.

4.4. Discussion

The experimental results carry several implications for the design and deployment of AI-driven resource allocation systems in smart cities. First, they demonstrate that high predictive accuracy is achievable without sacrificing fairness when explicit equity constraints are incorporated into the allocation pipeline. The marginal accuracy reduction of 0.8% relative to the unconstrained XGBoost model is a reasonable price for a 34.2% improvement in cross-district equity, particularly given that the affected communities are those most dependent on reliable public services.

Second, the qualitative usability evaluation indicates that SHAP-based explanations are practically meaningful to non-technical municipal stakeholders, supporting the vision of "human-in-the-loop" governance for algorithmic public systems. This finding aligns with the growing regulatory expectation, codified in the EU AI Act's "right to explanation" provisions, that consequential automated decisions must be accompanied by accessible justifications.

Third, the finding that Income Level ranks fifth rather than first in SHAP importance provides some reassurance that the model is not directly using a socioeconomic proxy as a primary allocation driver. However, we caution that fairness evaluation must extend beyond feature importance rankings: even features with low global importance may function as proxies in specific local decision contexts, underscoring the continued necessity of local explanation auditing at the individual decision level.

Regarding data transparency: the use of a synthetic dataset in this study is both a methodological strength and a current limitation. While it enables full reproducibility and eliminates privacy risks, the absence of a city-specific empirical ground truth means that

results should be interpreted as proof-of-concept evidence for the framework's effectiveness under controlled conditions. Future work should prioritize validation against real municipal datasets, which may exhibit distribution shifts, measurement errors, and governance constraints not captured by synthetic generation. The incorporation of causal fairness approaches—moving beyond correlational bias metrics—would additionally deepen the diagnostic power of the framework and improve decision robustness.

5. Conclusion

This paper has presented XAI-FA, a novel framework that integrates explainability and fairness-aware optimization for automated resource allocation in smart city environments. The framework combines a high-performance gradient-boosted decision tree classifier with SHAP global explanations, LIME local explanations, and Lagrangian-constrained fairness optimization. Evaluated on a purpose-built synthetic dataset comprising 120,000 records spanning five urban districts over a simulated three-year period, XAI-FA achieves 91.5% classification accuracy, a mean explainability score of 0.91, and a 34.2% reduction in inter-district fairness disparity relative to baseline black-box approaches.

These results contribute empirical evidence that the accuracy–fairness–explainability trade-off, often portrayed as a fundamental constraint, can be substantially mitigated through principled framework design. The proposed ECS and IDDI metrics offer reusable evaluation instruments for urban resource equity that may be adopted by both researchers and practitioners. The practical usability of the generated explanations, confirmed through a structured stakeholder evaluation, positions XAI-FA as a viable tool for transparent, accountable, and equitable governance of smart city infrastructure.

Future research will investigate: (i) extension of XAI-FA to dynamic, real-time allocation settings using online and adaptive learning techniques, enabling the system to respond to continuously evolving urban conditions; (ii) incorporation of causal fairness frameworks to move beyond associational bias detection and provide deeper insights into the root causes of disparate outcomes; (iii) validation across multiple cities with diverse socio-economic and infrastructural characteristics to strengthen generalizability and practical relevance; (iv) exploration of more granular allocation settings, including multi-class and continuous resource distribution; and (v) co-design of explanation interfaces with citizen communities through participatory design approaches, ensuring that transparency mechanisms genuinely serve public accountability rather than merely satisfying regulatory formalities.

Declaration on Generative AI

The authors used AI-assisted tools for language editing and text refinement during the preparation of this manuscript.

References

[1] Bibri, S.E. and Krogstie, J. (2017). Smart sustainable cities of the future: An extensive interdisciplinary literature review. Sustainable Cities and Society, 31, pp.183–212. https://doi.org/10.1016/j.scs.2017.02.016

[2] Brauneis, R. and Goodman, E.P. (2018). Algorithmic transparency for the smart city. Yale Journal of Law & Technology, 20(1), pp.103–176.

[3] Chen, T. and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, pp.785–794. https://doi.org/10.1145/2939672.2939785

[4] O'Neil, C. (2016). Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. New York: Crown Publishers.

[5] Eubanks, V. (2018). Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor. New York: St. Martin's Press.

[6] Adadi, A. and Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). IEEE Access, 6, pp.52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

[7] Lundberg, S.M. and Lee, S.I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems 30 (NeurIPS 2017). Long Beach, CA: Curran Associates, pp.4765–4774.

[8] Ribeiro, M.T., Singh, S. and Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, pp.1135–1144. https://doi.org/10.1145/2939672.2939778

[9] Arrieta, A.B. et al. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, pp.82–115. https://doi.org/10.1016/j.inffus.2019.12.012

[10] Guidotti, R. et al. (2018). A survey of methods for explaining black box models. ACM Computing Surveys, 51(5), Article 93, pp.1–42. https://doi.org/10.1145/3236009

[11] Jain, S. and Wallace, B.C. (2019). Attention is not Explanation. Proceedings of NAACL-HLT 2019. Minneapolis, MN: ACL, pp.3543–3556. https://doi.org/10.18653/v1/N19-1357

[12] Wachter, S., Mittelstadt, B. and Russell, C. (2018). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harvard Journal of Law & Technology, 31(2), pp.841–887.

[13] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), pp.206–215. https://doi.org/10.1038/s42256-019-0048-x

[14] Lu, R., Hong, S.H. and Zhang, X. (2018). A dynamic pricing demand response algorithm for smart grid: Reinforcement learning approach. Applied Energy, 220, pp.220–230. https://doi.org/10.1016/j.apenergy.2018.03.072

[15] Wang, Y. et al. (2015). Integrating demand response into energy management of isolated micro-grid with renewable energy generation. IET Renewable Power Generation, 9(5), pp.516–526. https://doi.org/10.1049/iet-rpg.2014.0172

[16] Mounce, S.R. et al. (2003). Sensor-fusion of hydraulic data for burst detection and location in a treated water distribution system. Information Fusion, 4(3), pp.217–229. https://doi.org/10.1016/S1566-2535(03)00034-4

[17] Mukhopadhyay, A. et al. (2022). A review of incident prediction, resource allocation, and dispatch models for emergency management. Accident Analysis & Prevention, 165, Article 106501. https://doi.org/10.1016/j.aap.2021.106501

[18] Caragliu, A. and Del Bo, C.F. (2022). Smart cities and urban inequality. Regional Studies, 56(7), pp.1097–1112. https://doi.org/10.1080/00343404.2021.1984421

[19] Mehrabi, N. et al. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6), Article 115, pp.1–35. https://doi.org/10.1145/3457607

[20] Dwork, C. et al. (2012). Fairness through awareness. Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS 2012). New York: ACM Press, pp.214–226. https://doi.org/10.1145/2090236.2090255

[21] Hardt, M., Price, E. and Srebro, N. (2016). Equality of opportunity in supervised learning. Advances in Neural Information Processing Systems 29 (NeurIPS 2016). Barcelona, Spain: Curran Associates, pp.3315–3323.

[22] Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big Data, 5(2), pp.153–163. https://doi.org/10.1089/big.2016.0047

[23] Kleinberg, J., Mullainathan, S. and Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. Proceedings of the 8th Innovations in Theoretical Computer Science Conference (ITCS 2017). Berkeley, CA: ACM Press, Article 43, pp.1–23. https://doi.org/10.4230/LIPIcs.ITCS.2017.43

[24] Yan, A. and Howe, B. (2020). Fairness-Aware Demand Prediction for New Mobility. Proceedings of the AAAI Conference on Artificial Intelligence, 34(1), pp.1079–1087. https://doi.org/10.1609/aaai.v34i01.5458

[25] Bellamy, R.K.E. et al. (2019). AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. IBM Journal of Research and Development, 63(4/5), pp.4:1–4:15. https://doi.org/10.1147/JRD.2019.2942287

[26] Lim, M.K., Li, Y., Wang, C. and Tseng, M.L. (2021). A literature review of blockchain technology applications in supply chains: A comprehensive analysis of themes, methodologies and industries. Computers & Industrial Engineering, 154, Article 107133. https://doi.org/10.1016/j.cie.2021.107133

[27] Bader, V. and Kaiser, S. (2019). Algorithmic decision-making? The user interface and its role for human involvement in decisions supported by artificial intelligence. Organization, 26(5), pp.655–672. https://doi.org/10.1177/1350508419856592