Explainable Deep Learning Framework for Predicting Urban Air Quality from Multisource Data Streams

Jack C. Bndersson
Department of Computer Science, University of Houston, Houston, TX, USA.
jacka@uh.edu

Abstract

Urban air quality prediction has emerged as a critical analytics problem demanding integration of heterogeneous data streams, high-fidelity modeling, and actionable explanations for decision support. This paper presents a comprehensive framework for an explainable deep learning system that ingests continuous multisource data, including regulatory monitoring stations, low-cost sensor networks, meteorological feeds, traffic telemetry, and satellite-derived aerosol optical depth, to produce high-resolution air quality forecasts. The architecture couples a spatiotemporal attention-based deep learning predictor with a suite of post-hoc and intrinsic interpretability modules, enabling the provision of human-readable rationales alongside each forecast. We discuss structural trade-offs in designing a system that balances prediction latency, model capacity, and faithfulness of explanations under resource constraints typical of municipal deployments. The governance architecture embedding the framework is examined with regard to data provenance, model versioning, and auditability, while robustness is analyzed under sensor drift, data missingness, and adversarial perturbation scenarios. Fairness considerations are foregrounded, highlighting how differential sensor density and model opacity can produce disparate impacts on marginalized communities. Policy implications for integrating explainable forecasts into public health alerting, traffic management, and urban planning are evaluated. The paper argues that high-accuracy prediction alone is insufficient; sustainable sociotechnical infrastructure requires built-in transparency mechanisms that align model reasoning with regulatory standards and community trust. A longitudinal deployment perspective is maintained, addressing operational sustainability, calibration decay, and institutional capacity-building.

Keywords

air quality prediction, explainable deep learning, multisource data streams, urban informatics, governance, fairness, spatiotemporal modeling.

1. Introduction

The deterioration of urban air quality constitutes one of the most pressing environmental health challenges in contemporary metropolitan regions. Exposure to elevated concentrations of particulate matter, nitrogen dioxide, ozone, and sulfur dioxide is consistently linked to respiratory and cardiovascular morbidity, premature mortality, and disproportionate burdens on economically disadvantaged populations. Consequently, accurate forecasting of pollutant concentrations at fine spatial and temporal granularity has become indispensable for early warning systems, dynamic traffic management, and long-term urban planning. In recent years, advances in deep learning have enabled significant improvements in predictive performance by learning complex nonlinear spatiotemporal dependencies from large-scale sensor data. Simultaneously, the proliferation of multisource data streams, ranging from official regulatory monitors to low-cost citizen-deployed sensors, meteorological reanalysis products, and satellite remote sensing, has vastly enlarged the feature space available for modeling. Yet the operational translation of these technical advances into trusted municipal decision support remains impeded by a critical gap: the opacity of high-capacity models renders their predictions uninterpretable to regulators, public health officials, and affected communities. Without transparent reasoning, even highly accurate forecasts fail to garner institutional adoption or enable effective contestability. The present work addresses this gap by proposing and examining a system-level architecture for an explainable deep learning framework dedicated to urban air quality prediction from multisource data streams. The discussion extends beyond model architecture to encompass data integration strategies, interpretability mechanisms, deployment infrastructures, and governance protocols, reflecting the position that explainability in this domain is fundamentally a sociotechnical property rather than a purely algorithmic one.

Urban air quality systems sit at the intersection of environmental science, data engineering, machine learning, and public policy, each introducing distinct constraints. Regulatory monitoring networks adhere to strict quality assurance protocols but offer sparse spatial coverage; low-cost sensors provide density but exhibit calibration drift and cross-sensitivity. Satellite observations deliver regional context but suffer from coarse temporal resolution and cloud-induced data gaps. Traffic and meteorological data stream in at high velocity with heterogeneous formats. A prediction framework must harmonize these streams while maintaining resilience to missingness, noise, and distribution shifts. The explainability imperative compounds these challenges, as explanation modules must handle time-varying feature importance, spatial heterogeneity, and correlation structures that defy simple attribution. Moreover, the explanations themselves must be aligned with the decision processes of end-users, whether they are environmental agency personnel assessing compliance with legal thresholds, urban planners siting new green infrastructure, or residents interpreting a health advisory. This paper contributes a holistic examination of the design space, analyzing architectural alternatives and their implications for robustness, fairness, and sustainability, and situating model-centric innovations within a broader governance framework.

2. Background and Related Work

The trajectory of air quality modeling has progressed from Gaussian dispersion models and chemical transport simulations to statistical land-use regression and, most recently, to data-driven machine learning approaches. Earlier works demonstrated that gradient boosting and random forests could effectively capture nonlinear relationships between land use, traffic, and pollutant levels, but their capacity to model temporal dynamics remained limited. The advent of deep recurrent architectures, particularly long short-term memory networks, provided a step change by learning temporal dependencies in time series of pollutant concentrations [1,2]. Convolutional neural networks were subsequently applied to model spatial dispersion patterns when monitoring stations were arranged on regular grids, with extensions to graph neural networks enabling irregular spatial topologies [3,4]. Attention mechanisms, first popularized in natural language processing, were adapted to air quality prediction to capture both short-term and seasonal dynamics, and to handle asynchronous multisource inputs [5,6]. These models have achieved state-of-the-art accuracy on public benchmarks, yet their black-box nature has spurred a parallel interest in interpretable machine learning.

Explainable artificial intelligence has developed along two broad paradigms: intrinsically interpretable models and post-hoc explanation methods. In air quality contexts, linear models and decision trees remain widespread in regulatory settings due to their transparency, though they sacrifice predictive fidelity. Generalized additive models with smooth splines preserve interpretability while allowing nonlinear effects, but struggle with high-dimensional interactions. Neural network interpretability has been advanced by techniques that compute input feature attributions via gradient-based methods, perturbation-based methods, or propagation rules tailored to recurrent or convolutional architectures [7,8]. Model-agnostic explanation techniques, such as those that perturb input features to approximate local decision boundaries [9], have been adapted to temporal settings by treating time steps as individual features or by generating temporal masks, though these adaptations entail computational cost and stability challenges. More recently, Shapley value-based explanations have been employed to distribute prediction outcomes among input features while satisfying desirable axiomatic properties [10]. In spatiotemporal air quality models, attention weights have been directly inspected as proxies for importance, though research cautions that attention may not reliably indicate feature attribution [11]. A growing body of work advocates for hybrid systems that combine a high-accuracy deep predictor with a separate explanation module optimized for different fidelity requirements and stakeholder audiences [12].

The intersection of multisource data fusion and explainability remains underexplored. Existing fusion architectures predominantly focus on improving predictive accuracy by concatenating embeddings from heterogeneous encoders, with limited consideration for how the contribution of each modality can be transparently quantified. While some studies report modality ablation analyses, such post-hoc evaluations do not constitute real-time explanation capable of supporting operational decisions. Furthermore, the deployment pipeline from research prototypes to sustained municipal services introduces challenges related to data governance, model updating, and algorithmic fairness that have only begun to be systematically addressed [13,14]. This paper extends the prior literature by articulating a full-stack perspective that integrates data fusion, deep learning architecture, explainability modules, and governance infrastructure within a unified framework.

3. Framework Architecture and System Design

The proposed framework is structured as a layered architecture comprising four principal subsystems: data ingestion and harmonization, spatiotemporal predictive modeling, explainability engine, and governance overlay. The data ingestion layer normalizes heterogeneous streams into a common spatiotemporal grid with configurable resolution, applying imputation strategies that are logged for downstream auditability. Sources include reference-grade monitors operated by environmental protection agencies, low-cost electrochemical and optical particle counter sensors maintained by municipalities or community groups, meteorological variables from numerical weather prediction outputs, real-time traffic flow and probe vehicle speed data, and satellite-derived aerosol optical depth and columnar trace gas retrievals. All incoming data are timestamped, geopositioned, and tagged with provenance metadata to enable lineage tracking. The layer implements versioned ingestion pipelines that can be redeployed upon schema changes without disrupting downstream model operations.

The predictive modeling layer accepts fused spatiotemporal tensors and generates forecasts for multiple pollutants at lead times ranging from one hour to seventy-two hours. The architectural backbone is a spatiotemporal attention-based encoder-decoder network with modality-specific encoders and a cross-modal transformer that learns inter-source dependencies dynamically. Meteorological and traffic streams are processed by separate temporal convolutional encoders, while satellite data pass through a vision transformer branch pretrained on large-scale earth observation datasets. The latent representations are aligned via a joint attention module before being passed to the decoder, which employs causal masking to preserve temporal precedence. The design intentionally avoids monolithic end-to-end training of all components from scratch, instead adopting a modular microservice pattern wherein each encoder can be independently retrained and versioned, facilitating maintenance and reducing technical debt. The trade-off incurred is a modest reduction in joint optimization capacity relative to a fully coupled architecture, but this is judged acceptable given the priority of long-term operational sustainability and explainability.

The explainability engine is co-located with the predictor and operates in both synchronous and asynchronous modes. In synchronous mode, each forecast that is served to an external consumer is accompanied by an explanation packet computed via a combination of integrated gradients for temporal inputs, layer-wise relevance propagation for spatial inputs, and Shapley value approximations for tabular features derived from point observations. The explanation packet includes local feature attributions, a simplified natural language summary for non-specialist users, and a set of counterfactual scenarios illustrating how changes in key drivers would alter the forecast. Asynchronous mode supports deeper retrospective analyses, such as global feature importance aggregation over seasons or pollution episodes, and subgroup analyses stratified by neighborhood income or demographic characteristics to probe fairness. The explainability engine is designed as a standalone service with a well-defined API, enabling future advancements in explanation algorithms to be integrated without retraining the core predictor.

4. Multisource Data Integration and Provenance

The complexity of multisource data integration arises not only from format heterogeneity but also from conflicting quality regimes and epistemic uncertainties. Reference monitors provide high-accuracy measurements at hourly resolution but are sparsely distributed, typically one per several tens of square kilometers. Low-cost sensors, by contrast, can be deployed in dense grids but exhibit drift, cross-sensitivity to humidity and temperature, and intermittent connectivity. Satellite retrievals offer wall-to-wall spatial coverage but at overpass revisit intervals that are insufficient for hourly forecasting and with retrieval biases under cloudy conditions. The framework reconciles these disparities through a data harmonization layer that performs spatial interpolation using Gaussian process regression informed by land-use covariates, temporal imputation via matrix factorization with missingness indicators, and bias correction of low-cost sensors against reference instruments using rolling calibration functions. All transformations are logged with metadata recording the imputation model version, the raw and adjusted values, and the identifiers of any sensors that supplied calibration data. This provenance trail is essential for regulatory audits and for debugging forecast anomalies, as it allows regulators to distinguish model errors introduced by data preprocessing from those inherent in the predictor.

The integration strategy intentionally preserves data source identifiers rather than collapsing all observations into an opaque aggregate. By maintaining source-specific channels, the explainability engine can attribute forecast contributions to satellite data versus ground sensors, thereby enhancing transparency. This design raises the question of weighting: a naive uniform weighting across sources would amplify noise from low-cost sensors, while hard filtering would discard valuable spatial information. The framework implements an adaptive reliability gating mechanism that learns, for each spatiotemporal cell, a dynamic weighting coefficient based on the recent performance of each source relative to the reference network. The coefficients themselves become an explainable output, visualized as trust maps that inform users where the model has higher confidence in its inputs. Such trust maps are communicated alongside forecasts in public-facing dashboards, fostering calibrated trust.

5. Deep Learning Model Design and Performance Trade-offs

The predictive core employs a hierarchical attention architecture that first computes self-attention within each data stream to capture modality-specific temporal dynamics, then applies cross-modal attention to fuse representations. The encoder for ground-based pollutant observations uses dilated causal convolutions to increase the receptive field without sacrificing chronological order, while meteorological encoder adopts a simpler Temporal Convolutional Network to satisfy latency constraints during inference. Cross-modal fusion is performed by a transformer with thirty-two attention heads, where the queries originate from the pollutant stream and keys and values from auxiliary streams, reflecting the causal structure that meteorology and traffic influence air quality but not vice versa on hourly timescales. The decoder generates multi-horizon forecasts autoregressively, with scheduled sampling to mitigate exposure bias. To enforce physical consistency, a soft penalty term is added to the loss function discouraging negative concentration predictions and extremely rapid concentration changes that violate atmospheric mixing constraints, though the penalty is tuned lightly to avoid suppressing genuine pollution spikes.

The tension between predictive accuracy and explainability fidelity is a central theme in architectural decisions. Highly overparameterized transformer models have demonstrated superior performance on benchmark datasets but produce attention distributions that are diffuse and difficult to interpret without additional stabilization techniques. The framework adopts a bottleneck fusion design, compressing each modality's representation to a lower dimension before cross-attention, which both regularizes the model and yields more focused attention weights that align better with known physical drivers, such as the dominant influence of traffic during morning rush hours. Ablation experiments conducted during development confirmed that the bottleneck reduces predictive error only marginally, while substantially improving the stability and actionability of explanation outputs. The design prioritizes monotonicity of key relationships where domain knowledge is strong: for instance, the dependence of nitrogen dioxide on traffic volume is constrained to be positive through architectural gating that prohibits negative gradients in that pathway, thus preventing counterintuitive explanations that would erode user trust.

Computational resource consumption is a further critical trade-off in municipal settings where the framework must operate on existing public cloud or on-premises infrastructure. The complete pipeline, including data preprocessing, inference, and explanation generation, is targeted to complete within five minutes for a seventy-two-hour forecast across a metropolitan area of ten million residents. To meet this latency budget, model distillation is employed, training a compact student network on the outputs of the full ensemble, and explanation algorithms are implemented with optimized sampling strategies that reduce the number of model evaluations required for Shapley value estimation. Dynamic batching and mixed-precision inference further reduce compute footprint, enabling deployment on GPU-equipped edge servers within environmental agencies.

6. Explainability Mechanisms and Stakeholder Alignment

Explainability in air quality systems must serve multiple stakeholder groups with distinct cognitive and operational needs. Environmental regulators require rigorous documentation of model inputs and decision factors to justify regulatory actions, such as issuing air quality alerts or establishing low-emission zones. Public health officials need interpretable risk communication that translates concentration forecasts into health advisories understandable by the general population. Community organizations and environmental justice advocates demand evidence that predictions do not systematically underestimate pollution in marginalized neighborhoods due to sensor underrepresentation. Satisfying these varied demands requires the framework to generate layered explanations that span algorithmic attribution, statistical evidence, and domain-grounded narratives.

The explainability engine produces a multi-tier explanation stack. The lowest tier comprises raw attribution vectors indicating the marginal contribution of each input feature to the forecast at each location and lead time. The intermediate tier aggregates attributions by data source, pollutant precursor, and spatial region, and performs statistical significance testing against historical baselines to suppress spurious attributions. The highest tier synthesizes these results into natural language templates tailored to each stakeholder group, with a configurable detail slider that allows users to drill down from a simple statement such as "Tomorrow's PM2.5 levels will be high primarily due to stagnant meteorological conditions and evening traffic emissions" to quantitative contributions with associated uncertainty bounds. The tiered design ensures that no single group is overwhelmed while preserving a path for deep investigation. A crucial element is the documentation of explanation uncertainty: the engine reports confidence intervals for attributions obtained through bootstrap resampling of the prediction ensemble, communicating that explanations themselves have error bars.

Counterfactual reasoning is integrated as a first-class explanation modality. Users can submit hypothetical interventions, such as "What would the nitrogen dioxide level be if traffic were reduced by thirty percent?" and receive model-based responses accompanied by the estimated attribution shift. This capability supports scenario planning for traffic management and urban design. The engine implements efficient counterfactual generation by performing a single forward pass with modified inputs through the same architecture, leveraging the model's learned conditional distributions rather than requiring full retraining. The counterfactuals are displayed alongside factual forecasts, with a visual distinction that prevents misinterpretation. To avoid misleading users, all counterfactuals carry a disclaimer that they assume other conditions remain constant, a condition that is explicitly verified by checking that the modified scenario lies within the convex hull of training data.

7. Deployment, Sustainability, and Governance

Deploying an explainable air quality prediction framework as a sustained municipal service necessitates institutional governance structures that extend far beyond model training. The framework is encapsulated in a containerized microservice architecture deployed via Kubernetes, with separate namespaces for data ingestion, prediction, explanation, and a public API gateway. Continuous integration and continuous delivery pipelines automate model retraining triggered either on a fixed weekly schedule or by data drift detectors that monitor the divergence between recent sensor readings and model predictions. Each retrained model is versioned and registered in a model registry, and predictions are always served with model version identifiers to ensure auditability. When agencies issue alerts based on model forecasts, they can trace back exactly which model version and input data produced the alert, enabling post-hoc forensic analysis.

Sustainability concerns encompass both environmental footprint and institutional capacity. The computational demands of deep learning raise legitimate questions about the carbon cost of continuous model retraining. The framework mitigates this by employing incremental learning strategies where only the final fusion layers are updated on new data, freezing the pretrained modality encoders for extended periods, and by scheduling training workloads during periods of high renewable energy availability on the cloud provider's grid when such information is accessible. More fundamentally, the framework is designed to be maintainable by municipal IT staff without specialized machine learning expertise. This is achieved through extensive automation, a well-defined model card system that documents intended use, performance characteristics, ethical considerations, and known limitations, and a user interface that exposes monitoring dashboards for data quality, prediction accuracy, and explanation consistency. The model card also provides a structured channel for agencies to report edge cases and anomalous explanations, feeding into an improvement lifecycle.

Governance must also address the tension between proprietary model components and public accountability. While agencies may procure the framework from a vendor, the core explanation methods and data provenance records must remain transparent and auditable by independent third parties. The architecture enforces this by requiring that all explanation APIs return results in a standardized open schema and that all data transformation code be open-sourced or at least exposed to regulatory auditors. Model explanations are logged alongside forecasts for a retention period defined by environmental information regulations, creating an evidentiary record that can be used in legal or administrative proceedings where air quality forecasts influenced decisions with economic or health consequences. This governance design treats explainability not merely as an interface feature but as a compliance requirement that shapes the entire lifecycle of the system.

8. Fairness, Robustness, and Policy Implications

Fairness in urban air quality prediction is inextricably linked to the spatial distribution of sensors and the demographics of affected populations. If low-income neighborhoods are underrepresented in the monitoring network, the data-driven model will exhibit higher prediction uncertainty in those areas, potentially leading to missed pollution episodes and delayed health interventions. The framework addresses this through several mechanisms. First, the data harmonization layer explicitly quantifies prediction uncertainty for each grid cell and propagates it through to the explanation outputs, so that users can observe where forecasts are less reliable. Second, fairness audits are conducted routinely by subgroup analysis, comparing prediction error metrics and explanation consistency across census tracts stratified by income, race, and health vulnerability indices. When disparities exceed predefined thresholds, the system issues an alert to the governance board and restricts usage in high-stakes decision contexts until corrective actions, such as deploying supplementary low-cost sensors or incorporating satellite data with appropriate bias correction, are implemented. Third, the framework includes an oversampling augmentation module that synthetically densifies training data in underrepresented regions using physics-informed generative models, thereby reducing model bias without requiring immediate physical sensor deployment. However, this approach carries risks of over-confident synthetic patterns if not carefully validated, so its outputs are clearly labeled as augmented and undergo additional expert review.

Robustness is tested against common operational failure modes: sensor drift, communication outages, and adversarial manipulation. Sensor drift is detected by comparing low-cost sensor readings against colocated reference instruments during periodic calibration campaigns, and by monitoring statistical divergences across the sensor fleet. The framework's reliability gating mechanism automatically down-weights drifting sensors, but it also maintains a record of sensors that were down-weighted, preventing silent exclusion. Communication outages leading to missing data are handled by multiple imputation chains, and the explainability engine annotates forecasts affected by substantial missingness with flags that convey reduced confidence. Adversarial robustness is considered in the context of deliberate manipulation, such as a malicious actor feeding falsified traffic data to avoid traffic restrictions. The architecture implements input validation ranges based on physical plausibility and historical extremes, and the anomaly detection module flags suspicious patterns for human review. While perfect robustness is unattainable, the design principle is that failures should be graceful and transparent.

The policy implications of explainable air quality frameworks are far-reaching. When regulators issue health advisories or activate traffic-limiting measures based on model predictions, the ability to articulate the rationale in terms of specific measured precursors and modeled atmospheric dynamics enhances both democratic legitimacy and legal defensibility. Municipalities can condition the procurement of air quality prediction services on compliance with explainability standards, much as they mandate open data formats. Standard-setting bodies can develop certification protocols that assess the faithfulness and stability of explanations, analogous to the validation of reference air quality instruments. At the same time, the availability of high-resolution, source-attributed predictions empowers community groups to hold polluters accountable by identifying spatial and temporal patterns that correlate with industrial activity or traffic routing decisions. The framework thus becomes not just a monitoring tool but an instrument of environmental governance, reshaping the distribution of information power between agencies, industry, and civil society. Ensuring that this power is exercised responsibly requires ongoing oversight, inclusive stakeholder engagement in system design, and continuous evaluation of both the model and its explanations against real-world outcomes.

9. Conclusion

This paper has presented a holistic framework for explainable deep learning prediction of urban air quality from multisource data streams, examining the architectural, operational, and governance dimensions that determine its effectiveness and trustworthiness. By integrating heterogeneous data through provenance-aware pipelines, coupling a spatiotemporal attention-based predictor with a multi-tier explainability engine, and embedding the system within a governance architecture that mandates transparency and auditability, the framework addresses the critical gap between predictive accuracy and actionable public decision-making. Structural trade-offs between model complexity and interpretability were analyzed, and the choice of modular, bottlenecked architectures was justified by the need for stable, attributable explanations over marginal accuracy gains. The discussion of fairness highlighted the disproportionate impacts that sensor distribution biases can produce and outlined mitigation strategies that combine technical augmentations with procedural fairness audits. Deployment sustainability was framed not only in terms of computational efficiency but also institutional maintainability, emphasizing model cards and automated retraining pipelines that reduce the burden on municipal agencies. The policy analysis situated explainability as a prerequisite for regulatory adoption and environmental justice, arguing that transparent models can transform air quality forecasting from a purely technocratic exercise into a participatory governance tool. Future work must extend longitudinal field validations that correlate the delivery of explanations with measurable changes in user trust and decision quality, and must address the challenges of scaling such frameworks to megacities across diverse regulatory regimes. The ultimate objective is a class of urban environmental intelligence systems that are not only accurate but also accountable.

References

1. Huang, C.-J., & Kuo, P.-H. (2018). A deep CNN-LSTM model for particulate matter (PM2.5) forecasting in smart cities. Sensors, 18(7), 2220.

2. Li, X., Peng, L., Yao, X., Cui, S., Hu, Y., You, C., & Chi, T. (2017). Long short-term memory neural network for air pollutant concentration predictions: Method development and evaluation. Environmental Pollution, 231, 997–1004.

3. Wang, J., & Song, G. (2018). A deep spatial-temporal ensemble model for air quality prediction. Neurocomputing, 314, 198–206.

4. Zhao, J., Deng, F., Cai, Y., & Chen, J. (2019). Long short-term memory-Fully connected (LSTM-FC) neural network for PM2.5 concentration prediction. Chemosphere, 220, 486–492.

5. Liang, Y., Ke, S., Zhang, J., Yi, X., & Zheng, Y. (2018). GeoMAN: Multi-level attention networks for geo-sensory time series prediction. Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, 3428–3434.

6. Zheng, Y., Yi, X., Li, M., Li, R., Shan, Z., Chang, E., & Li, T. (2015). Forecasting fine-grained air quality based on big data. Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2267–2276.

7. Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. Proceedings of the 34th International Conference on Machine Learning, 3319–3328.

8. Montavon, G., Binder, A., Lapuschkin, S., Samek, W., & Müller, K.-R. (2019). Layer-wise relevance propagation: An overview. In Explainable AI: Interpreting, Explaining and Visualizing Deep Learning (pp. 193–209). Springer.

9. Ribeiro, M. T., Singh, S., & 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, 1135–1144.

10. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765–4774.

11. Jain, S., & Wallace, B. C. (2019). Attention is not explanation. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 3543–3556.

12. Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82–115.

13. Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., ... & Schafer, B. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. Minds and Machines, 28(4), 689–707.

14. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... & Gebru, T. (2019). Model cards for model reporting. Proceedings of the Conference on Fairness, Accountability, and Transparency, 220–229.

15. Veeramachaneni, K., Vladen, I., & Re, C. (2016). Deep feature synthesis: Towards automating data science endeavors. Proceedings of the IEEE International Conference on Data Science and Advanced Analytics, 182–191.

16. Ma, J., Ding, Y., Cheng, J. C. P., Jiang, F., & Xu, Z. (2019). Soft detection of anomalous sensor data in IoT-based air quality monitoring systems. IEEE Internet of Things Journal, 6(6), 10776–10786.

17. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.

18. Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. Proceedings of the Conference on Fairness, Accountability, and Transparency, 59–68.

19. Zou, J., & Schiebinger, L. (2018). AI can be sexist and racist—it's time to make it fair. Nature, 559, 324–326.

20. Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. WIREs Data Mining and Knowledge Discovery, 9(4), e1312.

21. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press.

22. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998–6008.

23. Zhang, Y., & Wallace, B. (2017). A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. Proceedings of the Eighth International Joint Conference on Natural Language Processing, 253–263.

24. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794.