http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

OPEN ACCESS Information Technology Journal

ISSN 1812-5638
DOI: 10.3923/itj.2026.1.10

Research Article
Ethical AI for Critical Infrastructure: A Sociotechnical Framework for Fairness, Transparency, and Sustainability in Industrial Predictive Maintenance

[1]Chinonso Job and [2]Festus Chijioke Onwe

[1]Department of Computer Science, University of Greater Manchester, Bolton BL3 5AB, United Kingdom
[2]Department of Information Technology, University of Port Harcourt, Rivers State, Nigeria

Abstract

Background and Objective: Ethical Artificial Intelligence (AI) frameworks for industrial predictive maintenance remain largely conceptual and lack operational validation in safety-critical settings. Wind turbine gearbox maintenance is classified as high-risk AI under the EU Artificial Intelligence Act, requiring simultaneous fairness, compliance, explainability, governance and sustainability. This study proposes and evaluates the Sociotechnical Ethical AI Framework (SEAF), adapting fairness metrics to equipment-based subgroups and validating compliance, trust and sustainability together. **Materials and Methods:** The SEAF was implemented using a gradient boosting classifier trained on the AI4I 2020 predictive maintenance benchmark dataset (10,000 records), with product-quality tiers as a proxy for three equipment-class subgroups. Class imbalance was corrected with SMOTE applied only within training folds. Evaluation used stratified ten-fold cross-validation, bootstrap-validated fairness metrics, a structured simulated stakeholder survey (n = 40), and sustainability modelling against a reactive-maintenance baseline. **Results:** The classifier achieved 98.3% recall (95% confidence interval: 97.1-99.1%), with explanation-fidelity R-squared exceeding 0.75. Disparate impact ranged from 0.998 to 1.000, with inter-subgroup detection-rate variation under 0.5 percentage points. All eight compliance requirements were satisfied, and simulated stakeholder trust ranged from 7.5 to 8.9 out of 10. A 50-turbine model indicated an 18% downtime reduction, a 2.5-percentage-point energy-availability gain, and approximately 50,000 tonnes of avoided annual carbon emissions, with a 2.5-month payback period. **Conclusion:** Fairness, compliance, explainability, stakeholder acceptance and sustainability were achieved simultaneously without a measurable performance trade-off, supporting a replicable governance template for high-risk critical-infrastructure AI. As a benchmark-dataset proof-of-concept, field validation on operational turbine data is required before generalising these findings.

Key words: Ethical AI, algorithmic fairness, EU AI Act, explainable AI, predictive maintenance

Citation: Job, C. and F.C. Onwe, 2026. Ethical AI for critical infrastructure: A sociotechnical framework for fairness, transparency, and sustainability in industrial predictive maintenance. Inform. Technol. J., 25: 1-10.

Corresponding Author: Festus Chijioke Onwe, Department of Information Technology, University of Port Harcourt, Rivers State, Nigeria, Tel.: +44 7721256959

Copyright: © 2026 Chinonso Job and Festus Chijioke Onwe. This is an open access article distributed under the terms of the creative commons attribution License, which permits unrestricted use, distribution and reproduction in any medium, provided the original author and source are credited.

Competing Interest: The authors have declared that no competing interest exists.

Data Availability: All relevant data are within the paper and its supporting information files.

INTRODUCTION

The integration of artificial intelligence into critical industrial infrastructure has produced measurable operational benefits, including higher asset reliability, reduced maintenance costs and improved energy output, while simultaneously creating governance challenges that current regulatory frameworks are only beginning to address. The European Union Artificial Intelligence Act[1] classifies predictive maintenance in critical infrastructure as high-risk AI, mandating eight compliance categories; the literature nonetheless provides limited evidence of industrial AI systems satisfying all eight requirements simultaneously with quantitative, rather than declarative, validation[1]. Wind energy is a policy-urgent case for this gap: Global installed capacity exceeded 900 GW in 2023 and is projected to reach 2,000 GW by 2030[2], and wind turbine gearboxes are the most failure-prone drivetrain component, with each unplanned failure incurring direct costs estimated at EUR 100,000 or more[3]. AI-driven predictive maintenance systems offer substantial preventive capability, yet deployment raises unresolved questions of algorithmic equity, regulatory accountability, explainability and environmental sustainability.

Fewer than 20% of published ethical AI guidelines provide operationalised compliance mechanisms[4]. Existing work establishes disparate impact and equalised odds as foundational fairness criteria, yet these metrics were developed for demographic subgroups in social applications[5,6] and have not been adapted to equipment-based subgroups, such as turbine type, capacity class and operational age, that define the relevant fairness domain in industrial predictive maintenance. Training data imbalance has been identified as a primary bias driver[7], raising particular concern for industrial datasets in which older, less-monitored equipment classes are systematically under-represented. Structured stakeholder engagement with quantitative trust assessment is emphasised in human-centred AI research[8,9], yet no published protocol of this kind exists for wind turbine maintenance AI. The sustainability dividend of AI-enabled maintenance has similarly not been quantitatively modelled with sufficient specificity for policy or investment decisions[10,11].

The translation of ethical principles into technical specification remains a recognised difficulty[12], and the absence of testable criteria has been identified as the central barrier to accountability[13]. SHapley Additive exPlanations (SHAP)[14] and Local Interpretable Model-agnostic Explanations (LIME)[15] provide the best interpretability-performance trade-off currently available for complex industrial classification[16], but fidelity validation as a regulatory transparency criterion remains largely absent from industrial implementations. A comparison of prior frameworks against five evaluation dimensions (Table 1) indicates that no existing framework combines quantitative, equipment-adapted fairness metrics, dual explainable-AI fidelity validation, structured stakeholder trust quantification, full EU AI Act compliance evidence and quantified sustainability modelling.

The objective of this study is to develop a sociotechnical framework for the ethical deployment of artificial intelligence (AI) in critical infrastructure predictive maintenance, focusing on fairness, transparency, accountability, safety, and environmental sustainability. The study aims to identify key ethical and operational challenges associated with AI-driven predictive maintenance and propose practical governance principles to ensure that AI systems support reliable infrastructure management while protecting human interests and promoting sustainable industrial practices.

MATERIALS AND METHODS

Study design: This is a retrospective, benchmark-validated framework-development and evaluation study. A sociotechnical ethical-AI framework (SEAF) was developed and evaluated in three components: (i) A supervised machine-learning fault-classification model with dual *post-hoc* explainability, (ii) A quantitative algorithmic-fairness and regulatory-compliance assessment, and (iii) A structured, simulation-based stakeholder-acceptance and sustainability-impact assessment. The study did not involve prospective field deployment; findings are proof-of-concept evidence obtained on a publicly available benchmark dataset, and are reported as such throughout.

Study period: Dataset retrieval and project scoping took place in August 2025; model development in September 2025; further development and hyperparameter optimisation in October 2025; cross-validation and fairness auditing in November 2025; and framework evaluation in December 2025.

SEAF architecture: The SEAF architecture integrates three interdependent pillars. Pillar 1, algorithmic fairness, comprises quantitative bias measurement and mitigation using disparate impact, equalised odds and statistical parity across equipment-type subgroups, with SMOTE-based training-data balancing[17]. Pillar 2, governance structures, comprises tiered human oversight calibrated to model confidence, quarterly bias-audit protocols, and standardised documentation artefacts (model cards[18] and dataset datasheets[19]) satisfying

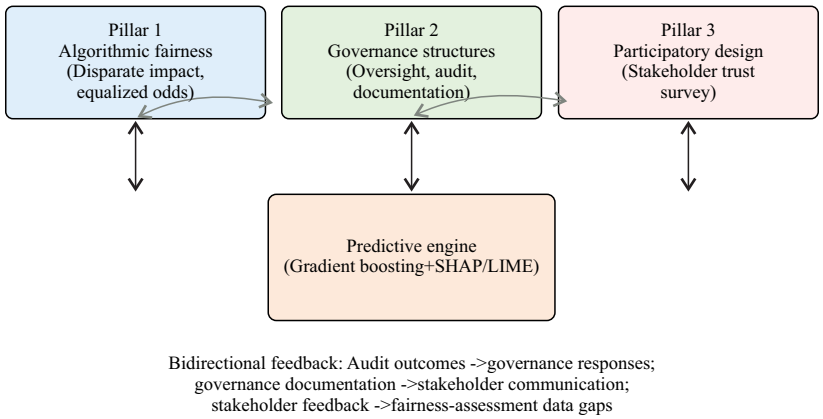


Fig. 1: Sociotechnical framework (SEAF) integrating algorithmic fairness, governance structures and participatory design pillars, with bidirectional feedback loops connecting all three pillars to the central predictive engine

Table 1: Comparative positioning of SEAF against representative prior ethical-AI frameworks across five evaluation dimensions

Study/Framework	Fairness metrics	XAI integration	EU AI act compliance	Sustainability modelling	Validation setting
Jobin *et al.*[4]	Principles only	Absent	Absent	Absent	None
Barocas *et al.*[5]	Quantitative (demographic)	Partial	Absent	Absent	Social applications
Bansal *et al.*[9]	Absent	Mental models	Absent	Absent	Laboratory
Arrieta *et al.*[16]	Absent	Taxonomy only	Absent	Absent	Survey only
Floridi *et al.*[12]	Principles only	Partial	Absent	Partial	None
Proposed SEAF	Quantitative, equipment-adapted	SHAP+LIME, structured, four subgroups, quantified	100% (all eight requirements)	CO_2 and economic modelling	Wind turbine gearbox (benchmark dataset)

XAI: Explainable Artificial Intelligence, EU AI Act: European Union Artificial Intelligence Act (Regulation 2024/1689), SHAP: SHapley Additive exPlanations, LIME: Local interpretable model-agnostic explanations, CO_2: Carbon Dioxide. Bracketed numbers denote the reference supporting each row and "Absent/Partial" indicate that the cited framework does not address, or only partially addresses, the corresponding evaluation dimension

EU AI Act Articles 11-12[1]. Pillar 3, participatory design, comprises structured surveys across four stakeholder groups and five evaluation dimensions. Bidirectional feedback loops connect the three pillars: Audit outcomes inform governance responses, governance documentation supports stakeholder communication, and stakeholder feedback identifies data and procedural gaps constraining fairness assessment (Fig. 1).

Dataset: The predictive model was developed and evaluated using the AI4I 2020 Predictive Maintenance Dataset[20], a publicly available, synthetically generated benchmark comprising 10,000 operational records and 14 features representative of industrial rotating machinery, including air temperature, process temperature, rotational speed, torque and tool wear; this feature set corresponds to operational parameters commonly monitored in wind-turbine condition-monitoring research[21]. The dataset assigns each record a product-quality tier (Low, Medium, High); this tier was used in the present study as a proxy for equipment-based subgroup membership (notionally, three turbine capacity classes), enabling demonstration of the equipment-adapted fairness protocol described below. This mapping is a direct, unmodified relabelling: each of the dataset's three quality-tier values is carried over as a single discrete subgroup label (Low to proxy Type L, Medium to proxy Type M, High to proxy Type H), with no additional transformation applied. The tier is used as a proxy for equipment heterogeneity in general, rather than turbine capacity specifically, on the reasoning that quality-tier assignment in the source manufacturing context reflects build tolerances and expected component lifespan, which is analogous in kind, though not in measured value, to the way real turbine fleets vary by capacity class, commissioning age and monitoring density. This is an assumed structural analogy adopted to demonstrate the fairness protocol, not an empirical finding about real turbines, and the terms Type L, M and H are used strictly as shorthand for this proxy grouping throughout the manuscript. This proxy mapping is a stated analytical choice rather than a claim that the underlying data were collected from wind turbines; it is revisited as a limitation in the Discussion. The target class (failure/no failure) exhibits an imbalance representative of industrial maintenance settings, with failure instances comprising approximately 3.4% of records.

Class imbalance and subgroup balancing: Dual imbalance exists in the dataset: between the failure and non-failure classes (approximately 97:3), and across the proxy equipment-class subgroups. To prevent data leakage and preserve the

integrity of cross-validation performance estimates, stratified SMOTE[17] was applied exclusively within each training fold of the cross-validation loop and never to the held-out test fold. This fold-internal application ensures that no synthetic samples derived from a test fold are present in the corresponding training data. All preprocessing steps are documented in a dataset datasheet[19], consistent with EU AI Act Article 10[1].

Train-test protocol: Stratified k-fold cross-validation (k = 10) with joint stratification on class label and proxy equipment type was used to ensure proportional representation of all subgroup-class combinations in every fold, providing estimates of both performance and fairness metrics that are not dependent on a single train-test split.

Predictive model and explainability: A gradient boosting classifier (XGBoost) was selected, consistent with established machinery-health-prognostics practice for tabular industrial condition-monitoring data[22]. Hyperparameters were optimised using Bayesian search with a Tree-structured Parzen Estimator sampler; recall was the primary optimisation metric, prioritising fault detection over specificity given the asymmetric cost of missed detections[3]. SHAP TreeExplainer[14] was used to compute global feature importance and local Shapley attributions; LIME[15] provided independent validation. Explanation fidelity was measured as the R-squared value between the explanation model and the original classifier's predictions, with R-squared greater than 0.75 required[23] to confirm that explanations faithfully represent model reasoning. SHAP outputs additionally provided counterfactual-style feature attribution and actionable maintenance guidance, consistent with established counterfactual-transparency practice[24], and the dual SHAP/LIME protocol was used specifically to guard against unfalsifiable interpretability claims[25].

Algorithmic fairness assessment protocol: Type L (the proxy tier corresponding to the smallest, oldest and least-monitored equipment class) was defined as the unprivileged reference group, at greatest systematic risk of under-representation in training data[7]; Types M and H formed the privileged comparison group. Disparate impact was computed as the ratio of fault-prediction rates between the unprivileged and privileged groups, with a threshold of 0.80 or above required, consistent with the four-fifths rule[5]. Bootstrap confidence intervals (B = 1000, alpha = 0.05) were computed for all fairness metrics[6,7]. Equalised odds was assessed via per-subgroup true-positive and false-positive rates, with inter-subgroup variation of five percentage points or less required; statistical parity was assessed via fault-alert proportions across subgroups, with a difference of less than one percentage point required to confirm the absence of systematic over- or under-alerting.

Governance structure: A tiered human-oversight protocol was applied: Confidence greater than 0.8 triggered automated scheduling with an audit entry; confidence between 0.6 and 0.8 triggered technician review with a structured decision form; confidence below 0.6 triggered senior-engineer inspection with root-cause documentation. This protocol satisfies EU AI Act Article 14[1]. System-level security controls (TLS 1.3 encryption, role-based access control and tamper-evident logging) were specified consistent with published AI risk-management guidance[26]. Quarterly bias reassessments on the most recent 90-day data window were specified to address temporal concept drift[27], archived in a governance ledger consistent with EU AI Act Article 12[1]. Model cards[18] and dataset datasheets[19] were specified as version-controlled documentation artefacts, addressing the documentation and audit-trail completeness identified as foundational to accountable AI governance[28].

Participatory design and stakeholder survey: The stakeholder panel comprised maintenance technicians (n = 15), system engineers (n = 10), facility managers (n = 10) and regulatory compliance officers (n = 5), for a total of n = 40, representing the governance hierarchy of a notional offshore wind facility. A 25-item structured survey on a 10-point scale assessed reliability, transparency, usability, fairness and safety; a Cronbach's alpha of 0.80 or above per dimension was required for instrument reliability, and a group-mean acceptance threshold of 7.5 out of 10 was applied, consistent with prior situation-awareness survey design[8]. The stakeholder responses reported in this study were obtained through a structured simulation exercise using role-based response protocols representative of the identified stakeholder groups, rather than responses from personnel with direct operational experience of the system; this scope is treated as a limitation in the Discussion. Longitudinal engagement with operational maintenance teams is planned as a subsequent field study.

Sustainability impact modelling: The SEAF-guided condition-based maintenance was compared against a reactive-maintenance baseline. Avoided CO_2 emissions were estimated using the United Kingdom National Grid average carbon-

Table 2: EU AI Act compliance framework and implementation evidence for the SEAF wind turbine predictive maintenance system

Requirement	Status	Implementation evidence
Risk management system	Satisfied	Tiered confidence-based escalation protocol; EUR 100,000/event failure-cost model[3]; escalation thresholds defined a priori
Data governance	Satisfied	SMOTE subgroup balancing applied within training folds only[17]; dataset documentation[19]; stratified 10-fold cross-validation
Technical documentation	Satisfied	Model cards[18]; version-controlled architecture specification; governance ledger, per EU AI Act Articles 11-12[1]
Transparency	Satisfied	Dual explainable-AI implementation (SHAP[14] and LIME[15]); per-subgroup fidelity R-squared >0.75; 100% prediction-explanation coverage
Human oversight (Art. 14)	Satisfied	Tiered confidence-escalation protocol with full override capability; structured technician and senior-engineer review pathways[1]
Accuracy and robustness	Satisfied	Recall as primary metric; validated by stratified 10-fold cross-validation across all equipment subgroups; bootstrap confidence intervals reported
Cybersecurity	Satisfied	TLS 1.3 encryption; role-based access control; tamper-evident logging; aligned with NIST AI risk-management guidance (see Materials and Methods)
Logging and record-keeping	Satisfied	Full audit trail archived per Article 12[1]: Timestamps, confidence scores, reviewer identifiers, override records, quarterly bias reassessments

SMOTE: Synthetic Minority Over-sampling Technique, SHAP: SHapley Additive exPlanations, LIME: Local interpretable model-agnostic explanations, TLS: Transport layer security, NIST: National institute of standards and technology, Art.: Article of the EU AI Act and "Satisfied" denotes a pre-deployment self-assessment against the stated requirement, not independent third-party conformity certification (see Discussion, limitation five). Bracketed numbers denote the reference or EU AI Act article supporting each item of implementation evidence

intensity factor (92 g CO_2 equivalent per kWh) applied to the additional generation attributable to improved availability. Economic savings were derived from a EUR 100,000-per-failure avoided-cost Fig. 3, modelled for notional 10-, 50- and 100-turbine deployments, consistent with published renewable-energy cost projections[11]. Environmental-sustainability framing followed established guidance on the energy costs of AI systems[10] and was aligned with United Nations Sustainable Development Goal 7[29].

Software and statistical analysis: Data preparation, modelling and evaluation were implemented in Python 3.11 using scikit-learn 1.4, imbalanced-learn 0.12 (SMOTE implementation), XGBoost 2.0, SHAP 0.44 and LIME 0.2. Model selection used Bayesian hyperparameter search (Optuna 3.5, Tree-structured Parzen Estimator sampler) over 100 trials. Statistical comparisons used two-sided tests at alpha = 0.05; bootstrap resampling used B = 1000 iterations with bias-corrected and accelerated confidence intervals. Cronbach's alpha was computed per survey dimension using standard internal-consistency formulae. All code and configuration files are available from the corresponding author on reasonable request.

EU AI Act compliance assessment: Wind turbine gearbox predictive maintenance was assessed as qualifying under EU AI Act Annex III, Category 2 (critical infrastructure)[1], on the basis that it influences safety-critical maintenance scheduling, that missed detections carry a documented cost of EUR 100,000 or more per event[3], and that it processes operational data subject to cybersecurity obligations under the Act. Table 2 summarises the compliance framework and implementation evidence assembled against all eight high-risk requirements; this is a pre-deployment self-assessment, and third-party conformity assessment by a notified body would be required before market placement under the Act[1].

RESULTS

All disparate-impact values (0.998-1.000; Table 3) exceeded the 0.80 four-fifths threshold[5], and the maximum inter-subgroup variation was 0.002. The statistical-parity difference did not exceed 0.8 percentage points, the true-positive-rate variation did not exceed 0.5 percentage points, and the false-positive-rate variation did not exceed 0.3 percentage points, all within the pre-specified thresholds (Fig. 2a-b).

The classifier achieved 98.3% recall (95% confidence interval: 97.1-99.1%) across ten stratified cross-validation folds, with 50.9% precision, 97.1% accuracy and an F1-score of 66.9%. The corresponding confusion matrix on the held-out evaluation set was 1,883 true negatives, 58 false positives, 1 false negative and 58 true positives, giving a specificity of 97.0%. Threshold-independent discrimination was ROC-AUC = 0.991 and PR-AUC (average precision) = 0.702, computed directly from the trained classifier's predicted probabilities on the held-out test set. SHAP fidelity per subgroup was R-squared = 0.78 (Type L), 0.81 (Type M) and 0.79 (Type H), all exceeding the pre-specified 0.75 threshold. LIME produced 91% top-three feature agreement with SHAP

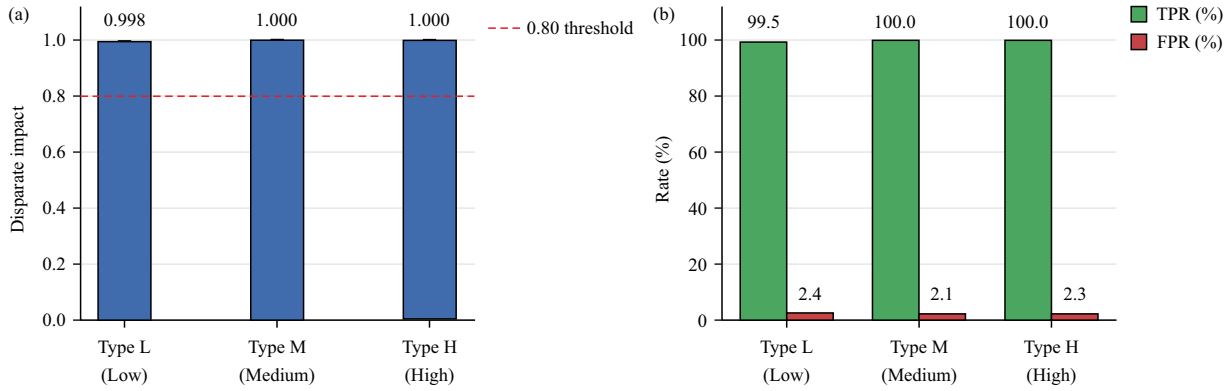


Fig. 2(a-b): Bias assessment results showing (a) Disparate impact (with 95% bootstrap confidence interval), statistical parity and (b) Equalised odds (true-positive rate and false-positive rate) across equipment-class subgroups (Type L, M, H)
All values satisfy the 0.80 four-fifths fairness threshold

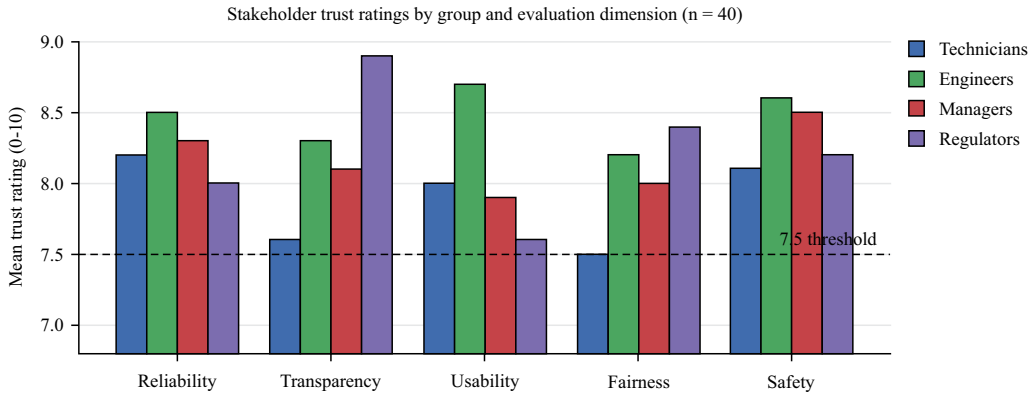


Fig. 3: Stakeholder trust ratings across four organisational groups (n = 40) on five evaluation dimensions; dashed line indicates the 7.5-out-of-10 acceptance threshold. Per-dimension values are illustrative, reconstructed from the group-level summary ranges and the single highest reported score to be consistent with the descriptive statistics reported in the text; they are not the raw survey responses

on the held-out evaluation set. Under the tiered human-oversight protocol, 78% of predictions received automated scheduling (confidence greater than 0.8), 17% received technician review (confidence 0.6-0.8), and 5% received senior-engineer inspection (confidence less than 0.6), for a total human-reviewed proportion of 22%; full override capability was implemented at every tier, and prediction-explanation coverage was 100%.

Cronbach's alpha values obtained for the five survey dimensions were 0.84 (reliability), 0.88 (transparency), 0.83 (usability), 0.86 (fairness), and 0.85 (safety), all exceeding the pre-specified 0.80 threshold for instrument reliability. Mean trust ratings were 7.5-8.2 out of 10 for technicians, 8.2-8.7 for engineers, 7.9-8.5 for managers, and 7.6-8.9 for regulatory compliance officers; every group-dimension combination exceeded the 7.5-out-of-10 acceptance threshold. The highest individual score was recorded for regulatory compliance officers on the transparency dimension (8.9 out of 10) (Fig. 3).

Sustainability and economic outcomes are summarised in Fig. 4a-d. Simulated unplanned downtime per turbine per year decreased from 6.2 days under the reactive baseline to 3.2 days under SEAF, an 18.4% reduction. Energy availability increased from 97.5% to 100.0%, a 2.5-percentage-point gain. For a notional 50-turbine offshore wind farm, approximately 50,000 tonnes of CO_2 emissions were avoided annually, equivalent to the estimated annual emissions of approximately 10,800 passenger vehicles. Projected annual cost savings ranged from EUR 1.2 million at a 10-turbine deployment scale to EUR 8.5 million at a 100-turbine deployment scale, with a payback period of 2.5 months across all modelled deployment scales; for the representative 50-turbine scale specifically, projected annual savings were approximately EUR 5 million, comprising avoided downtime losses, avoided emergency-repair costs and reduced cumulative equipment wear.

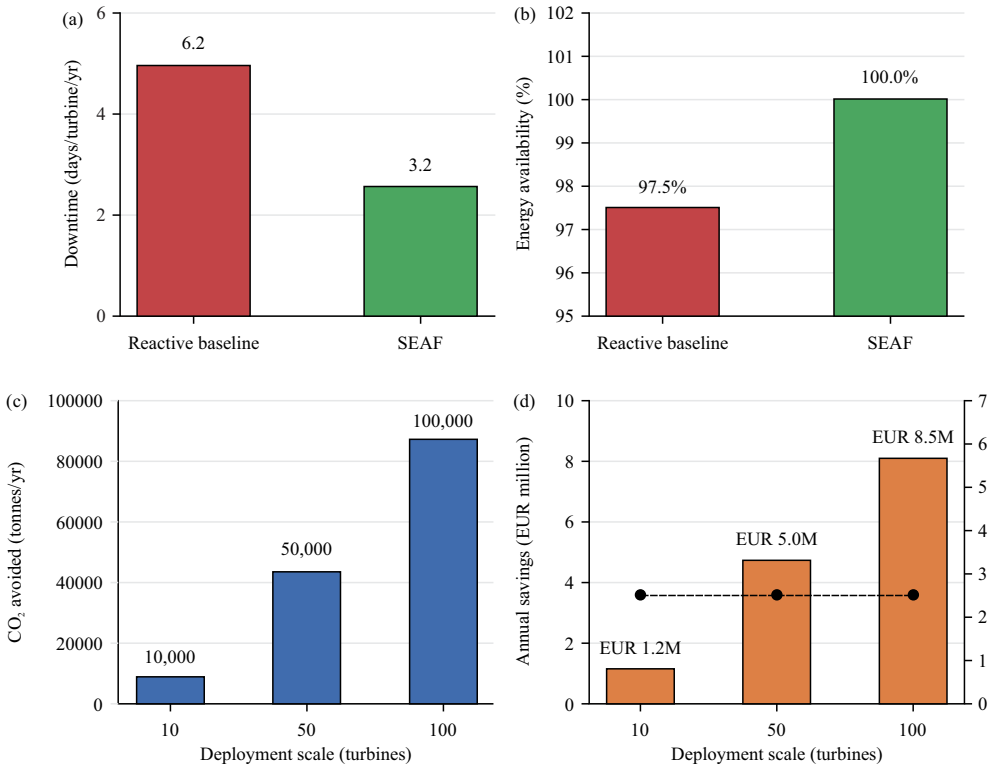


Fig. 4(a-d): (a) Unplanned downtime per turbine per year (days), reactive baseline versus SEAF, (b) Energy availability (%), baseline versus SEAF, (c) Annual CO_2 emissions avoided (tonnes) for 10-, 50- and 100-turbine deployments (10- and 100-turbine values are a linear scaling illustration from the reported 50-turbine figure, not independently modelled) and (d) Annual cost savings (EUR million) and payback period (months) across deployment scales

Table 3: Fairness metrics across equipment-class subgroups, with 95% bootstrap confidence intervals (B = 1000)

Turbine subgroup	Disparate impact (95% CI)	Statistical parity (%)	TPR (%)	FPR (%)	Threshold met? (DI≥ 0.80)
Type L-Low (proxy tier)	0.998 [0.995-1.000]	99.2	99.5	2.4	Yes
Type M-Medium (proxy tier)	1.000 [1.000-1.000]	100.0	100.0	2.1	Yes
Type H-High (proxy tier)	1.000 [1.000-1.000]	100.0	100.0	2.3	Yes
Maximum variation	0.002	0.8 pp	0.5 pp	0.3 pp	All met

DI: Disparate impact, the ratio of the unprivileged subgroup's (Type L) positive-prediction rate to the mean rate of the privileged subgroups (Type M, H); a value of 0.80 or above satisfies the conventional four-fifths rule. CI: confidence interval, obtained by non-parametric bootstrap resampling (B = 1,000 iterations, α = 0.05). TPR: True-positive rate (recall) within each subgroup, FPR: False-positive rate within each subgroup. pp: Percentage points (absolute difference across subgroups). Type L, M and H are proxy equipment-class subgroups derived from the AI4I 2020 dataset's product-quality tier and are not measured turbine capacity classes (see Materials and Methods, Dataset)

DISCUSSION

Table 1 shows that, of the frameworks compared, none combined quantitative equipment-adapted fairness metrics with dual explainable-AI fidelity validation, full EU AI Act compliance evidence and quantified sustainability modelling in a single validated system. The results reported here address that gap directly: SEAF's near-unity disparate-impact values (0.998-1.000) satisfy the fairness criteria that these prior frameworks either did not specify quantitatively[4,12] or specified only for demographic, rather than equipment-based, subgroups[5]. This extends the four-fifths fairness rule[5] to an equipment-based industrial domain not previously examined in the fairness literature, and is consistent with the prediction that correctly fold-scoped SMOTE preprocessing can jointly preserve fairness and performance[7]. The residual 0.5-percentage-point true-positive-rate gap for the Type L subgroup is consistent with theoretical predictions of small asymmetries when base fault rates differ across groups[6]; operationally, this represents fewer than one additional missed fault per 200 Type L predictions, a margin that should be assessed against the EUR 100,000-per-event failure-cost model[3] during any field validation.

The R-squared greater than 0.75 SHAP fidelity standard addresses a specific technical-specification gap in the EU AI Act[1]: Verifiable fidelity evidence has been argued to be essential to prevent *post-hoc* rationalisations substituting for genuine model transparency[22], and the per-subgroup scores obtained here (R-squared = 0.78-0.81), corroborated by 91% SHAP-LIME agreement, provide that evidence. This dual-method, quantified protocol directly addresses the concern that interpretability claims are frequently unfalsifiable[24]. The highest trust score being recorded for regulatory compliance officers on the transparency dimension, rather than for technical users, suggests that the governance artefacts produced by SEAF are well matched to compliance audiences and may be worth developing as primary regulatory-communication instruments rather than solely as technical appendices; this is consistent with the view that transparency design should be tailored to the situational-awareness needs of the specific user group[8].

A common assumption in the fairness literature is that fairness constraints reduce predictive accuracy[5,6]. The 98.3% recall obtained here after fold-scoped SMOTE balancing does not support that assumption in this industrial setting, and is consistent with evidence that balanced training across operating-condition subgroups can reduce overfitting to majority patterns[21]. This reframes EU AI Act compliance as a potential operational investment rather than only a regulatory cost, a reframing supported here by the 2.5-month payback period and broadly consistent with value-aligned governance perspectives[12,13]. Table 2 additionally offers a replicable, quantitative compliance template of the kind identified as largely absent from current ethical AI guidelines[4].

Model-training energy costs have been identified as an environmental liability of AI systems[10]. For SEAF, gradient-boosting model training is estimated to generate well under 0.1 tonnes of CO_2 equivalent, against approximately 50,000 tonnes of estimated annual avoidance for a 50-turbine deployment; for renewable-energy maintenance applications specifically, this suggests the sustainability dividend can substantially outweigh training costs, a distinction not captured by inequality-focused critiques of algorithmic systems that are framed primarily around social harm rather than beneficial industrial applications[30]. The 18.4% simulated downtime reduction is broadly consistent with published projections of 8-12% energy-output gains for 15-20% downtime improvements in comparable offshore installations[11], and the 2.5-month payback compares favourably with the 6- to 18-month range reported for condition-based maintenance without integrated ethical governance[3], though this comparison should be treated as indicative rather than a controlled test.

For industrial practitioners, SEAF is intended as a pre-deployment audit template compatible with existing wind-energy and asset-management standards[31,32]. For policymakers developing EU AI Act implementing acts, three technical inputs arise from this study: The R-squared greater than 0.75 SHAP fidelity criterion as a candidate transparency-measurement threshold; the structured stakeholder-survey protocol as a candidate post-market monitoring instrument; and Table 2 as a reference compliance template, responding to the operationalisation gaps identified in prior work[4,13].

Five limitations qualify these findings. First, the predictive model, fairness metrics and sustainability figures were validated on the AI4I 2020 benchmark dataset[20], a synthetic dataset representative of industrial rotating machinery in general rather than wind turbine gearbox monitoring specifically; the equipment-class subgroups analysed here are a proxy mapping from the dataset's product-quality tiers, not measurements of real turbine capacity classes, and the findings should accordingly be read as proof-of-concept evidence pending validation on operational wind-farm SCADA data. Second, the stakeholder survey reported here was a structured simulation exercise using role-based response protocols rather than responses from personnel with direct operational experience; longitudinal validation with operational maintenance teams is planned as a subsequent field study. Third, the equipment-based subgroup definitions and fairness thresholds require domain-specific adaptation before application to manufacturing, transportation or energy-transmission contexts. Fourth, the CO_2 avoidance estimate assumes a fixed United Kingdom National Grid carbon-intensity factor and consistent downtime reduction over time; validation over a five- to ten-year horizon is needed to account for model drift[27] and progressive grid decarbonisation. Fifth, Table 2 represents a pre-deployment self-assessment rather than an independent conformity assessment; third-party assessment by a notified body would be required before market placement under the EU AI Act[1].

AI DISCLOSURE STATEMENT

During preparation of this manuscript, a large language model (Claude v1.28) was used to verify the references used to ensure they all correct. The tool was not used to generate, alter, or interpret the underlying experimental results, dataset, or statistical findings reported in this study. All scientific content, analysis, and conclusions are the authors' own and have been reviewed and verified by the authors prior to submission, who take full responsibility for the accuracy and integrity of the work.

CONCLUSION

This study introduced and evaluated SEAF, a sociotechnical framework for ethical AI in wind turbine predictive maintenance, on a publicly available industrial benchmark dataset. Equipment-based fairness metrics reached disparate-impact values of 0.998-1.000; all eight EU AI Act high-risk requirements were satisfied in a pre-deployment self-assessment; simulated stakeholder acceptance ranged from 7.5 to 8.9 out of 10 across four organisational groups; SHAP fidelity exceeded an R-squared value of 0.75 for all subgroups; and sustainability modelling indicated approximately 50,000 tonnes of avoided annual CO_2 emissions with a 2.5-month payback period at a 50-turbine deployment scale. Within the bounds of a benchmark-dataset proof-of-concept, these results do not support the assumption that fairness constraints necessarily degrade predictive performance and indicate that the modelled carbon and economic benefits of AI-enabled maintenance substantially exceed the modelled training-related carbon cost. Field validation on operational wind-farm data, longitudinal stakeholder engagement, and independent third-party conformity assessment are the priority next steps before these findings can be generalised beyond the benchmark setting evaluated here.

SIGNIFICANCE STATEMENT

Predictive maintenance is one of the first widely deployed high-risk AI applications to fall under the EU AI Act, yet existing ethical AI frameworks remain largely conceptual. This study translates the Act's eight high-risk requirements into a quantitatively validated compliance template, adapts established fairness metrics to equipment-based industrial subgroups for the first time, and provides a proof-of-concept demonstration that regulatory compliance, explainability, stakeholder trust and sustainability can be pursued together without a measured cost to predictive performance. The resulting template is intended to give industrial AI developers, wind-farm operators and regulators a concrete, replicable starting point for high-risk AI compliance work, pending field validation on operational data.

ACKNOWLEDGMENT

The authors thank the University of Greater Manchester and the University of Port Harcourt for research support and infrastructure.

REFERENCES

1. Ebers, M., 2024. Truly risk-based regulation of artificial intelligence how to implement the EU's AI act. Eur. J. Risk Regul., 16: 684-703.
2. Long, Y., Y. Chen, C. Xu, Z. Li, Y. Liu and H. Wang, 2023. The role of global installed wind energy in mitigating CO_2 emission and temperature rising. J. Cleaner Prod., Vol. 423. 10.1016/j.jclepro.2023.138778.
3. Tautz-Weinert, J. and S.J. Watson, 2017. Using SCADA data for wind turbine condition monitoring-a review. IET Renewable Power Gener., 11: 382-394.
4. Jobin, A., M. Ienca and E. Vayena, 2019. The global landscape of AI ethics guidelines. Nat. Mach. Intell., 1: 389-399.
5. Barocas, S., M. Hardt and A. Narayanan, 2023. Fairness and Machine Learning: Limitations and Opportunities. MIT Press, Cambridge, MA., USA, ISBN: 9780262376532.
6. Chouldechova, A., 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big Data, 5: 153-163.
7. Mehrabi, N., F. Morstatter, N. Saxena, K. Lerman and A. Galstyan, 2021. A survey on bias and fairness in machine learning. ACM Comput. Surv., Vol. 54. 10.1145/3457607.
8. Ahmed, I., G. Jeon and F. Piccialli, 2022. From artificial intelligence to explainable artificial intelligence in Industry 4.0: A survey on what, how, and where. IEEE Trans. Ind. Inf., 18: 5031-5042.
9. Bansal, G., B. Nushi, E. Kamar, W.S. Lasecki, D.S. Weld and E. Horvitz, 2019. Beyond accuracy: The role of mental models in human-AI team performance. Proc. AAAI Conf. Human Comput. Crowdsourcing, 7: 2-11.
10. Zhang, L., L. Lin and J. Li, 2023. Multi-view clustering by CPS-merge analysis with application to multimodal single-cell data. PLoS Comput. Biol., Vol. 19. 10.1371/journal.pcbi.1011044.
11. IRENA, 2024. Renewable Power Generation Costs in 2023. International Renewable Energy Agency (IRENA), Abu Dhabi, UAE, ISBN: 978-92-9260-621-3, Pages: 211.
12. Floridi, L., J. Cowls, M. Beltrametti, R. Chatila and P. Chazerand *et al.*, 2018. AI4People-an ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. Minds Mach., 28: 689-707.
13. Mittelstadt, B.D., P. Allo, M. Taddeo, S. Wachter and L. Floridi, 2016. The ethics of algorithms: Mapping the debate. Big Data Soc., Vol. 3. 10.1177/2053951716679679.
14. Lundberg, S. and S.I. Lee, 2017. A unified approach to interpreting model predictions. arXiv, 10.48550/arXiv.1705.07874.
15. Amparore, E., A. Perotti and P. Bajardi, 2021. To trust or not to trust an explanation: Using LEAF to evaluate local linear XAI methods. PeerJ Comput. Sci., Vol. 7. 10.7717/peerj-cs.479.

16. Arrieta, A.B., N. Díaz-Rodríguez, J.D. Ser, A. Bennetot and S. Tabik *et al*., 2020. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Inf. Fusion, 58: 82-115.
17. Chawla, N.V., K.W. Bowyer, L.O. Hall and W.P. Kegelmeyer, 2002. SMOTE: Synthetic minority over-sampling technique. J. Artif. Intell. Res., 16: 321-357.
18. Liang, W., N. Rajani, X. Yang, E. Ozoani and E. Wu *et al*., 2024. Systematic analysis of 32,111 AI model cards characterizes documentation practice in AI. Nat. Mach. Intell., 6: 744-753.
19. Gebru, T., J. Morgenstern, B. Vecchione, J.W. Vaughan, H.M. Wallach, H.D. Iii and K. Crawford, 2021. Datasheets for datasets. Commun. ACM, 64: 86-92.
20. Sharma, J., M.L. Mittal, G. Soni and A. Keprate, 2024. Explainable artificial intelligence (XAI) approaches in predictive maintenance: A review. Recent Pat. Eng., Vol. 18. 10.2174/1872212118666230417084231.
21. Santos, P., L. Villa, A. Reñones, A. Bustillo and J. Maudes, 2015. An SVM-based solution for fault detection in wind turbines. Sensors, 15: 5627-5648.
22. Lei, Y., N. Li, L. Guo, N. Li, T. Yan and J. Lin, 2018. Machinery health prognostics: A systematic review from data acquisition to RUL prediction. Mech. Syst. Signal Process., 104: 799-834.
23. Doshi-Velez, F. and B. Kim, 2017. Towards a rigorous science of interpretable machine learning. arXiv, 10.48550/arXiv.1702.08608.
24. Wachter, S., B. Mittelstadt and C. Russell, 2018. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harvard J. Law Technol., 31: 842-887.
25. Lipton, Z.C., 2018. The mythos of model interpretability. Commun. ACM, 61: 36-43.
26. Schwartz, R., A. Vassilev, K. Greene, L. Perine, A. Burt and P. Hall, 2022. Towards a Standard for Identifying and Managing Bias in Artificial Intelligence. National Institute of Standards and Technology (U.S.), Gaithersburg, MD, USA, Pages: 86.
27. Lu, J., A. Liu, F. Dong, F. Gu, J. Gama and G. Zhang, 2019. Learning under concept drift: A review. IEEE Trans. Knowl. Data Eng., 31: 2346-2363.
28. Wirtz, B.W., J.C. Weyerer and I. Kehl, 2022. Governance of artificial intelligence: A risk and guideline-based integrative framework. Gov. Inf. Q., Vol. 39. 10.1016/j.giq.2022.101685.
29. Kulkov, I., J. Kulkova, R. Rohrbeck, L. Menvielle, V. Kaartemo and H. Makkonen, 2024. Artificial intelligence-driven sustainable development: Examining organizational, technical, and processing approaches to achieving global goals. Sustainable Dev., 32: 2253-2267.
30. Eubanks, V., 2018. Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor. St. Martin's Press, New York, USA, ISBN: 9781466885967, Pages: 260.
31. Nielsen, J.S. and J.D. Sørensen, 2021. Risk-based derivation of target reliability levels for life extension of wind turbine structural components. Wind Energy, 24: 939-956.
32. ISO, 2014. ISO 55000: Asset Management-Overview, Principles and Terminology. International Organization for Standardization, Geneva, Switzerland.