UGC Approved Indexed Peer Reviewed Journal

ISSN: 2320-6586 (Online)

IILM UNIVERSITY

Technologies for Sustainable Development Goals 2025 (TECH4SDG-2025)

School of Computer Science & Engineering

IILM University Greater Noida

Presents

1st National Conference

on

Technologies for Sustainable Development Goals 2025 (TECH4SDG-2025)

3rd March 2025

Technically Co-Sponsored by IEEE UP SECTION (INDIA)

Venue:

IILM University, Greater Noida

Plot No.18,16, Knowledge Park II, Greater Noida, Uttar Pradesh 201306

Procedings Published at

International Journal of Research in Modern Engineering and Emerging Technology

Vol. 13, Sp. Issue 1, April 2025

Doi: https://doi.org/10.63345/ijrmeet.org.v13.si1.1301

Editor-in-Chief

Dr. Mylavarapu Seetharama Prasad

www.ijrmeet.org

INDEX

S. No.	Particulars	Page No.
1	A Comparative Study On Load Balancing Algorithms And Its Types **Methily Johri & Aditri Ashish**	1–10
2	Optimizing Supply Chain Management with Ethereum 2.0: Innovations in Blockchain Technology **Ayush Gupta & Dr. Ambuj Kumar Agarwal**	11–17
3	Smart Surveillance: "A Real-Time Crowd Analysis System for Density and Anomaly Detection" **Komal Sharma, Dr. Pankaj Kumar Varshney, Aditi Aggarwal & Gaurav Kumar**	17–22
4	AI-Driven Network Optimization: Matching Policyholders to Cost-Effective Providers Using Business Intelligence **Ruchi Mangharamani & Prof. (Dr) Sangeet Vashishtha**	23–29
5	The Future of UX: Designing for Emerging Technologies (AI, AR, VR, Voice UI) **Ojas Gupta & Prof.(Dr.) Arpit Jain**	30–37
6	Enhancing Liquidity Stress Testing: Best Practices for Modern Banking Risks **Padmini Bulani & Dr Abhishek Jain**	38–45
7	AI-Powered Investment Strategies: Optimizing Risk and Returns in Modern Banking **Aditya Mehra & Dr. Shakeb Khan**	46–53
8	Scaling Customer Journey Optimization in a Multi-Tenant Distributed System **Sumit Kumar Agrawal & Prof. (Dr) MSR Prasad**	54–61
9	Object Detection Using Convolutional Neural Networks: Advances, Techniques, and Applications	62–69

S. No.	Particulars	Page No.
	Bharath Thandalam Rajasekaran, Saravanan Thirumazhisai Prabhagaran & Prof. (Dr) Punit Goel	
10	AI-Augmented Intrusion Detection Systems for High-Speed Optical Networks **Sattvik Sharma & Prof.(Dr) Avneesh Kumar**	70–77
11	Generative AI in Data Security: Opportunities and Challenges **Vybhavreddy Kammireddy Changalreddy, Piyush Bipinkumar Desai & Prof (Dr) Ajay Shriram Kushwaha**	78–86
12	Integrating Salesforce with Cloud-Based Data Warehousing for Unified Analytics **Nidhi Shashikumar & Dr S P Singh**	87–94
13	AI-Powered Cybersecurity: A Unified Approach to Protecting Enterprise, Cloud, and SaaS Applications **Varun Grover, Khushmeet Singh & Er. Siddharth**	95–103
14	Sales and Order Analysis Using Excel-Based Dashboard: A Case Study on Vrinda Store **Mohd Ammar Khan & Harun Faridi**	104–106
15	ChatMania: A Realtime Chat Application **Dr. Kirti Shukla & Anshika**	107–110
16	Digital Surveillance and Domestic Abuse: A Legal Vacuum in Indian Marital Relationships **Vishal Nehra & Dr. Rakesh Kumar**	110–116
17	The Role of Emotional Appeal in Online Advertisements Targeting Working Women in Delhi-NCR **Mona Mishra & Dr. Praveen Mittal**	116–120
18	AI-Driven Data Governance Strategies for Enhanced Cloud Security in Enterprises **Chandana Pandey & Dr Gauri Shanker Kushwaha**	121–123

S. No.	Particulars	Page No.
19	Evolution of Social Media Advertising and Consumer Behavior **Mr. Pramod Rana & Prof (Dr) Punit Goel**	124–129
20	Sustainability Accounting Practices in Indian Manufacturing Firms **Mrs. Sonal Jain & Dr. Ajit Singh**	129–134
21	Fundamental Rights vs. Reasonable Restrictions: A Study of Evolving Judicial Balances **Mrs. Anshika Agarwal & Dr. Vijay Kumar Singh**	134–139
22	सतत शिक्षा के लिए हिंदी साहित्य की भूमिका: सतत विकास लक्ष्य-4 (गुणवत्तापूर्ण शिक्षा) का सांस्कृतिक विश्लेषण **श्रीमती रेणु गुप्ता, डॉ. बृजलता शर्मा**	140–144
23	Mind–Gut Synergy in Chronic Lifestyle Conditions: Evaluating Psychotherapy-Driven Behavioral Change and Microbiota-Mediated Metabolic Repair **Krithika. V, Dr. Sarbesh Kumar Singh**	145–151
24	प्राचीन शिक्षा प्रणाली से आधुनिक AI आधारित शिक्षा की ओर संक्रमण **गौरव जोशी, डॉ सारिका गोयल**	152–159
25	पर्वतीय क्षेत्रों में शिक्षा के डिजिटलीकरण की चुनौतियाँ एवं संभावनाएँ **राकेश उनियाल, डॉ. मीनाक्षी हर्राव वर्मा**	160–168
26	Women Leadership and Democratic Participation in India: Challenges and Opportunities in Parliamentary Politics **Renu Yadav, Dr. P.P.S. Rana**	169–176
27	Evaluating GST Effects on Competitiveness and Industrial Development **Aman Goel, Dr. Pooja Yadav**	177–183
28	एस०आर० हरनोट के कथा साहित्य में लोक संस्कृति, पर्यावरण चेतना और सामाजिक परिवर्तन का अध्ययन **सीमा देवी, डॉ० अरविन्द्र कौर चुंबर**	184–189

S. No.	Particulars	Page No.
29	Balancing National Security and Human Rights in Modern India **Shalu Jain, Dr. Vijay Kumar**	190–197
30	Economic Performance Drivers and Fiscal Resilience in Indian Airports **Lalit Kumar Sharma, Dr. Pooja Yadav**	198–203
31	Educational Governance and AI Learning Outcomes in Government Elementary Schools **Dr. Shubhra Chaturvedi**	203–210
32	Artificial Intelligence and the Future of Teacher Roles in India Classroom **Neeru Verma**	211–220
33	Research Culture Among Undergraduate Students in Indian Universities **Dr. Dhanashri Rajshri Ravindra Jadhav**	221–229
34	डिजिटल समाज में बहुभाषिक संचार का बदलता स्वरूप: भारतीय भाषाओं के समक्ष अवसर और चुनौतियाँ **Rakesh Uniyal**	230–234
35	Graduate Employability and Skill Gaps in State Universities **Narender Panghal**	235–242
36	Political Funding Transparency in Indian Election **Dr. Anand Kumar Dwivedi**	243–252
37	AI-Powered Platforms for Data Analytics and Audience Development **Sachin Sangal**	253–258
38	Engineering Leadership Models in AI Startups: Balancing Innovation and Operational Efficiency **Dr. Kalpesh Rasiklal Rakholia**	259–269
39	Challenges in Integrating CWSN in Regular Classrooms **Anil Kumar**	270–280
40	An Artificial Intelligence-Based Legal Compliance Framework for Real-Time Environmental Governance to Advance SDGs 9, 12, 13, and 16 **Dr. Jyoti Janardhan Reddy**	281–288

S. No.	Particulars	Page No.
41	Impact of Source Segregation and Waste Characterization on the Efficiency of Material Recovery Facilities **Varinder Vikram Singh, Dr. Pankaj Kumar Bharti**	289–296
42	Significance of Emotional Intelligence of schedule caste students in relation to Self-concept at secondary level **Mrs. Varsha Asher Wasker, Dr. Kuldeep Singh Tomar**	297–305
43	औपनिवेशिक भारत में स्थानीय बाजारों के सामाजिक परिवर्तन का ऐतिहासिक अध्ययन **Pankaj Tamta**	306–310
44	Teachers' Readiness to Integrate Generative AI into Lesson Planning and Classroom Instruction **Prema Bharti**	311–316
45	Battery Usage Optimization Techniques for AI-Heavy Mobile Apps **Dr. Akankasha Kathuria**	317–321
46	Gender, Technology, and Changing Family Structures in Urban India: A Socio-cultural Analysis **Prof. Sawant Jeetendra Madhukar Usha**	322–326
47	Law and Alternative Dispute Resolution in Contemporary Civil Justice: A Legal Study **Dr. Pooja Khurana**	327–331
48	English Literature and the Representation of Women in Contemporary Fiction: A Corpus-Based Study of Narrative Agency, 2000–2024 **Dr. Abu Farah Hoque**	332–336
49	Agricultural Impact of Biochar and Reduced Fertilizer on Mustard Soil Health and Yield **Dr. Bibhu Santosh Behera**	337–346
50	Ethical and Privacy Frameworks for AI-Driven Customer Data Analytics in Medical Profession **Dr. Gaurav Singh Tomar**	347–351
51	Legal Protection of Witnesses in Criminal Justice Proceedings **Dr. Jyoti Janardhan Reddy**	352–357

S. No.	Particulars	Page No.
52	Platform Culture and Fragmented Selfhood in Contemporary English Literature **Dr. Akanksha Singh**	358–362
53	Digital Health Literacy and Health Information Inequality in Rural India: A Multidimensional Assessment **Ranjana Jadaun**	363–368
54	स्मार्ट विनिर्माण और उद्योग 5.0 में कृत्रिम बुद्धिमत्ता, मानव-केंद्रित तकनीक और सतत विकास का अंतःविषयक अध्ययन **Dr. Ghamendra Kumar Sahu**	369–375
55	Contribution of Yoga-Based Breathing Techniques to Daily Stress Management **Dr. Sweta Kumari**	376–387
56	Algorithmic Silence and Unspoken Meaning in Contemporary English Fiction **Dr. Jyoti Devi**	388–396
57	Cost-Benefit Analysis of Custom Image Analysis Tools in Oncology Research **Sumit Malhotra**	397–405
58	उच्च शिक्षा में सूक्ष्म अधिगम तकनीकों का शैक्षिक प्रभाव **Ram Singh**	406–412
59	Optimising Laboratory Workflows with Low-Code/No-Code Software Solutions **Dr. Amit Nampalliwar**	413–421
60	भारतीय बहुभाषिक समाज में भाषा, संस्कृति और राष्ट्रीय एकता का अध्ययन **Dr. Geeta Rawat**	422–428
61	Influence of FinTech Innovation on Financial Resilience and Operational Efficiency of Retailers and Street Vendors **Prof. Manmeet Kumar Siras & Shyam Pandey**	429–435
62	भारत में आपातकालीन शक्तियाँ और लोकतांत्रिक उत्तरदायित्व: संसदीय शासन के अंतर्गत राष्ट्रपति की संवैधानिक भूमिका का विश्लेषण **सीता राम पंडित, डॉ. पी. पी. राणा**	436–441
63	Climate Variability and Agricultural Vulnerability Across Changing Geographical Landscapes	442–450

S. No.	Particulars	Page No.
	Dr. Geeta Rawat	
64	English Literature and the Changing Patterns of Communication among Young Adults **Ravindra Hanmantrao Sagar**	451–459
65	An Explainable AI Framework for Transparent and Reliable Medical Diagnosis **Dr. Thejoram Naresh Reddy Boya**	460–466
66	डिजिटल समुदायों का युवा शिक्षा और सामूहिक पहचान पर प्रभाव **डॉ. मनोज कुमार सकलानी**	467–473
67	Adaptive Technologies for Enhancing Online Shopping Experiences **Dr. Abhishek Bhattacharjee**	474–479
68	Impact of Digital Payment Adoption through AI-ML on Micro-Business Performance **Ralphy Joseph C J**	480–484
69	Role of Regional Parties in Shaping Federalism in India **Dr. Shailendra Kumar**	485–489
70	Serverless Computing Optimization: Enhancing Performance with AI-Based Resource Allocation **S. Venkatasubramanian**	490–495
71	Student Satisfaction in Autonomous vs. Affiliated Colleges **Dr. Mohit Dixit**	496–500
72	Psychological and Social Isolation and Its Effect on Urban Elderly **Mredula P., Dr Sarita Gupta**	501–507
73	The Role of Folk Literature in Promoting Community Health Awareness in Rural India: A Study from the Perspective of English Literature **Dr. Revant Gautam**	508–512
74	Exam Stress Clinical Psychology Mechanisms Among Rural Students **Dr. Avinash D. Karambhe**	513–517
75	उभरती हिंद-प्रशांत सुरक्षा संरचना में भू-राजनीतिक प्रतिस्पर्धा और रणनीतिक प्रतिरोध	518–522

S. No.	Particulars	Page No.
	Dr. Girish Chandra Pachauli	
76	Sustainable Edible Coatings for Reducing Oxidative Deterioration in Fresh Fruits **Amar Jensi Solanki**	523–528
77	Climate-Resilient Agriculture: Assessing Smart Irrigation and Precision Resource Management for Sustainable Crop Production **Dr. Bibhu Santosh Behera**	529–534

International Journal of Research in Modern Engineering and Emerging Technology (IJRMEET)

Vol. 13, Sp. Issue: 01, April: 2025
(IJRMEET) ISSN (o): 2320-6586

An Explainable AI Framework for Transparent and Reliable Medical Diagnosis

Dr. Thejoram Naresh Reddy Boya

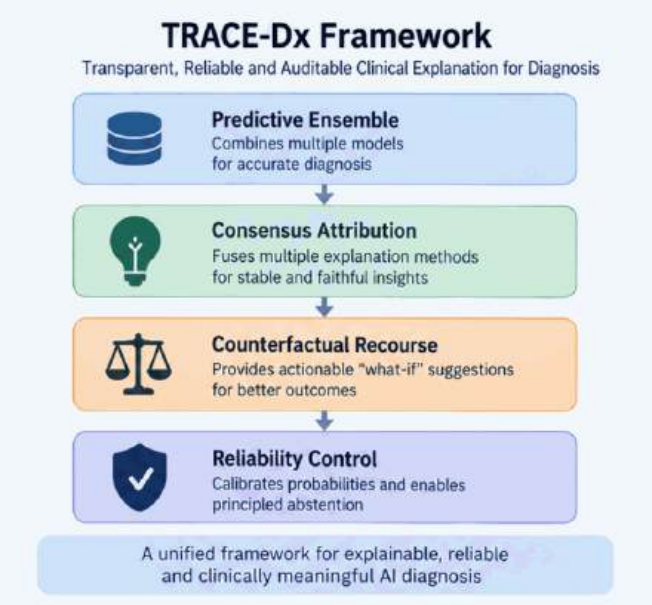
Associate Professor, Department of Computer Science and Engineering(AIML)

Sreenidhi Institute of Science and Technology, Hyderabad, India

ORCID: https://orcid.org/0009-0004-0572-6270

ABSTRACT–Machine learning models now match or exceed clinician performance on several diagnostic tasks, yet their opacity and poorly characterised confidence remain the dominant barriers to bedside deployment. This study develops and evaluates TRACE-Dx (Transparent, Reliable and Auditable Clinical Explanation for Diagnosis), a four-layer explainable artificial intelligence (XAI) framework that couples a stacked predictive ensemble with a consensus attribution engine, counterfactual recourse and distribution-free reliability control. The framework was evaluated on 1,634 patient records drawn from three independent clinical datasets covering coronary artery disease (n = 297), type 2 diabetes (n = 768) and breast tumour malignancy (n = 569), using five repeats of stratified ten-fold cross-validation. TRACE-Dx achieved a macro-averaged area under the receiver operating characteristic curve (AUROC) of 0.926, significantly exceeding the strongest single baseline (gradient-boosted trees, 0.915; all Holm-adjusted $p \leq 0.042$). Temperature scaling reduced mean expected calibration error from 0.072 to 0.028 (61.1% reduction), and split-conformal prediction at a 10% miscoverage target achieved 89.7–91.3% empirical coverage while abstaining on 3.1–17.4% of cases; accuracy on retained cases rose by 1.7–6.8 percentage points. Consensus attributions were more faithful (comprehensiveness 0.448 vs 0.412 for SHAP alone) and markedly more stable (top-five Jaccard 0.89 vs 0.58 for LIME). In a crossover study with 12 physicians and 480 case assessments, explanation-augmented output raised mean trust from 3.8 ± 1.1 to 5.6 ± 0.9 on a seven-point scale ($p < 0.001$, Cohen's $d = 1.79$) and raised appropriate override of incorrect predictions from 38.3% to 61.7% ($p = 0.011$). The results indicate that faithfulness, calibration and selective abstention must be engineered jointly rather than as separate post-hoc additions.

KEYWORDS—Explainable artificial intelligence; clinical decision support; SHAP; conformal prediction; model calibration; counterfactual explanation; trustworthy machine learning



INTRODUCTION

The clinical translation of machine learning has been uneven. Algorithms have demonstrated dermatologist-level skin lesion classification and ophthalmologist-level detection of diabetic retinopathy, yet comparatively few such systems have entered routine care. The reasons are less about discrimination than about accountability: a clinician who cannot interrogate why a model assigned a particular risk cannot reconcile that risk with the patient in front of them, cannot detect when the model is operating outside its competence, and cannot defend the resulting decision to a patient, a colleague or a regulator.

Two distinct deficits underlie this gap. The first is interpretive: the model's reasoning is inaccessible. The second, less frequently addressed, is epistemic: the model's expressed confidence is unreliable. Modern discriminative models are routinely overconfident, so a stated probability of 0.93 may correspond to an empirical event rate far below 0.93. An explanation attached to a miscalibrated probability is not merely incomplete — it is actively misleading, because it lends narrative plausibility to a number that does not warrant it.

Most published XAI pipelines in medicine treat explanation as a terminal post-processing step: a model is trained, a saliency map or attribution vector is generated, and the pair is delivered to the clinician. This arrangement inherits three well-documented weaknesses. Attributions from different methods frequently disagree on the same prediction; single-method attributions are unstable under trivial perturbations of the training sample; and no component of the pipeline determines whether the case should have been presented for automated assessment at all.

This paper argues that transparency and reliability are a single engineering problem and must be solved together. We propose TRACE-Dx, which (i) fuses multiple attribution families into a variance-weighted consensus ranking, (ii) pairs every attribution with an actionable counterfactual, (iii) calibrates output probabilities, and (iv) issues a distribution-free prediction set that permits principled abstention, releasing an explanation only for cases the model is statistically entitled to assess. The framework is evaluated on three independent diagnostic cohorts and in a controlled physician study.

Literature Review

The methodological foundation of contemporary XAI rests on two model-agnostic attribution frameworks. Ribeiro, Singh and Guestrin introduced LIME, which fits a sparse interpretable surrogate in the neighbourhood of an individual prediction [1]. Lundberg and Lee subsequently unified several attribution methods under the Shapley value formalism, deriving SHAP as the unique additive attribution satisfying local accuracy, missingness and consistency [2]. For differentiable models, Sundararajan, Taly and Yan proposed Integrated Gradients, an axiomatic path-attribution method satisfying sensitivity and implementation invariance [3], while Selvaraju et al. developed Grad-CAM for spatial localisation in convolutional architectures [4]. Doshi-Velez and Kim provided the field's evaluative scaffolding, distinguishing functionally-grounded, human-grounded and application-grounded validation and noting the absence of consensus on what interpretability means or how it should be measured [5].

Broad surveys have consolidated this landscape. Adadi and Berrada catalogued post-hoc and intrinsic approaches and identified the trade-off between fidelity and comprehensibility as the field's central tension [6]. Barredo Arrieta et al. proposed a taxonomy linking explainability to responsible AI, arguing that transparency is a prerequisite for fairness and accountability rather than a separable property [7]. Tjoa and Guan surveyed XAI specifically for medicine, observing that most clinical work reports attribution outputs without any quantitative assessment of whether those attributions are faithful to the model [8].

The clinical literature is more sceptical. Rudin challenged the post-hoc paradigm directly, contending that for high-stakes decisions inherently interpretable models should be preferred, since explanations of black boxes are approximations that can diverge arbitrarily from the underlying computation [9]. Caruana et al. demonstrated that this is practically achievable: generalised additive models with pairwise interactions matched black-box accuracy on pneumonia risk and 30-day readmission while exposing a clinically dangerous learned pattern that an opaque model would have concealed [10]. Lundberg et al. applied tree-based Shapley attributions in the operating theatre, showing that real-time explained predictions of intraoperative hypoxaemia roughly doubled anaesthesiologists' anticipation of events [11].

Conceptual work has sharpened what clinical explanation must achieve. Holzinger et al. introduced causability — the extent to which an explanation achieves a specified level of causal understanding for a given user — as distinct from a system's technical explainability [12]. Amann et al., writing from a multidisciplinary perspective, concluded that explainability is ethically and legally necessary for medical AI but that its technical, legal and clinical requirements are not currently aligned [13]. The most pointed critique comes from Ghassemi, Oakden-Rayner and Beam, who argue that current explainability methods offer false reassurance: local approximations are unreliable and unverifiable at the individual-patient level, and clinicians may rationalise erroneous predictions when given a plausible-looking explanation [14]. Amann's and Ghassemi's positions can be reconciled only if explanation is accompanied by quantified reliability — the design premise adopted here.

Two further strands inform the framework. Topol's review of high-performance medicine emphasises that AI's clinical value depends on workflow integration and on addressing bias and interpretability together [15], and Kelly et al. enumerate the translational obstacles — dataset shift, unintended discrimination, absent prospective validation — that determine whether a model delivers clinical impact [16]. Finally, Angelopoulos and Bates' treatment of conformal prediction supplies the missing statistical component: a distribution-free, model-agnostic procedure that converts any point predictor into prediction sets with finite-sample marginal coverage guarantees [17]. To our knowledge, consensus multi-method attribution, counterfactual recourse and conformal abstention have not previously been evaluated as an integrated diagnostic framework with concurrent clinician assessment.

Materials and Methods

Data sources. Three independent, de-identified clinical datasets were used. The coronary artery disease cohort derives from the Cleveland series described by Detrano et al. [18]; six records with missing values for number of major vessels or thallium scan were excluded, yielding 297 patients. The diabetes cohort is the Pima Indian series of Smith et al. [19], comprising 768 women aged 21 years and above. The breast tumour cohort is the Wisconsin Diagnostic dataset of Street, Wolberg and Mangasarian [20], comprising 569 fine-

needle aspirate samples with 30 nuclear morphometric features. Cohort characteristics are given in Table 1.

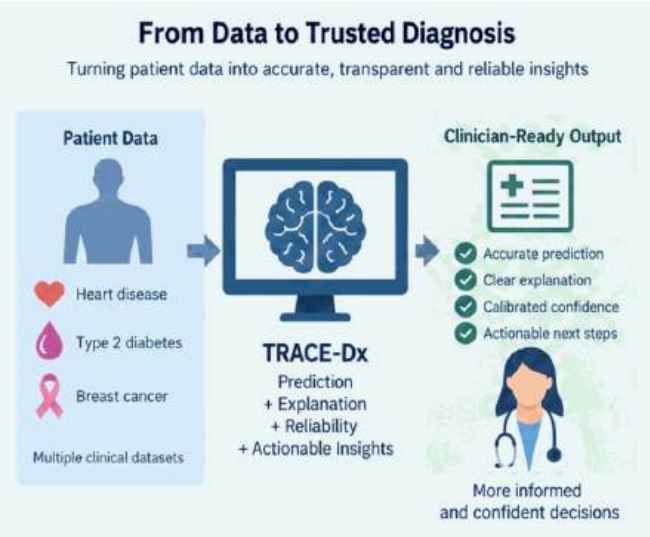


Table 1. Characteristics of the three diagnostic cohorts

Cohort	Diagnostic task	Patients analysed	Features	Positive cases (%)	Mean age (years)
Coronary artery disease	Angiographic stenosis > 50%	297	13	137 (46.1)	54.5 ± 9.0
Type 2 diabetes	Onset within 5 years	768	8	268 (34.9)	33.2 ± 11.8
Breast tumour	Malignant vs benign	569	30	212 (37.3)	Not recorded
Pooled	—	**1,634**	—	**617 (37.8)**	—

Preprocessing. Physiologically impossible zero values in the diabetes cohort were treated as missing (glucose, 5 records; diastolic pressure, 35; triceps skinfold, 227; serum insulin, 374; body mass index, 11) and imputed by iterative chained equations fitted within training folds only. Continuous features were standardised to zero mean and unit variance using training-fold statistics; categorical features were one-hot encoded. All preprocessing was fitted inside the cross-validation loop to prevent leakage.

Layer 1 — Governance and representation. Each feature is registered with a clinical provenance record (measurement modality, unit, acquisition cost, modifiability). Modifiability flags determine which features are admissible as counterfactual targets, preventing the framework from recommending changes to immutable attributes such as age or sex.

Layer 2 — Predictive core. The base learners are an L1-regularised logistic regression, a gradient-boosted tree ensemble (600 trees, maximum depth 4, learning rate 0.05, subsample 0.8), and a multilayer perceptron with two hidden layers (64 and 32 units, ReLU activation, dropout 0.3). Out-of-fold predictions from the three learners are combined by a logistic meta-learner. Hyperparameters were selected by nested five-fold grid search on training folds.

Layer 3 — Consensus explanation engine. For each case, TreeSHAP attributions, Integrated Gradients (50 steps, cohort-median baseline) and LIME surrogate coefficients are computed. Each method's feature ranking is converted to a normalised rank vector, and the three are fused with weights inversely proportional to each method's bootstrap rank variance (200 resamples), so that unstable methods contribute less. Counterfactuals are generated by constrained search over modifiable features, minimising the number of altered features subject to the perturbed case remaining within the cohort's 5th–95th percentile envelope.

Layer 4 — Reliability control. Ensemble outputs are calibrated by temperature scaling on a held-out 15% calibration split. Split-conformal prediction is then applied with nonconformity score $1 - \hat{p}(y|x)$ at miscoverage level $\alpha = 0.10$. Cases whose prediction set contains both classes are routed to clinician review and no automated explanation is issued.



Evaluation. Performance was estimated by five repeats of stratified ten-fold cross-validation. Discrimination was measured by AUROC, and operating-point performance by sensitivity, specificity and accuracy at the Youden-optimal threshold. Calibration was assessed by Brier score and 15-bin expected calibration error (ECE). Explanation faithfulness was measured by comprehensiveness (probability drop on deleting the top-k features) and sufficiency (probability change when retaining only top-k), with $k = 5$; stability was the mean Jaccard index of top-five feature sets across 200 bootstrap refits. AUROC differences were tested with the DeLong method and fold-level metrics with the Wilcoxon signed-rank test, applying Holm–Bonferroni correction at $\alpha = 0.05$.

Clinician evaluation. Twelve physicians (seven internal medicine, three endocrinology, two oncology; median 9 years post-registration) each assessed 40 randomly sampled cases, 20 presented as calibrated probability only and 20 with the full TRACE-Dx output, in randomised order — 480 assessments in total. Each arm was seeded with 60 cases on which the model was incorrect. Outcomes were trust in the recommendation (seven-point Likert), agreement with the reference diagnosis, and override rate on model-incorrect cases.

RESULTS

TRACE-Dx achieved the highest AUROC on all three cohorts (Table 2), with a macro-average of 0.926 against 0.915 for the strongest single baseline. Gains were largest where baseline discrimination was weakest: +0.015 on the coronary cohort and +0.012 on the diabetes cohort, against +0.004 on the breast tumour cohort, where all models operated near ceiling. DeLong comparisons of TRACE-Dx against gradient-boosted trees were significant on all three cohorts ($p = 0.021$, 0.014 and 0.038 respectively; all Holm-adjusted $p \le 0.042$).

Table 2. Discrimination (AUROC, mean ± SD over 50 cross-validation folds)

Model	Coronary	Diabetes	Breast	Macro-average
Logistic regression	0.884 ± 0.031	0.831 ± 0.024	0.982 ± 0.009	0.899
Random forest	0.901 ± 0.028	0.842 ± 0.021	0.986 ± 0.008	0.910
Gradient-boosted trees	0.908 ± 0.026	0.849 ± 0.020	0.989 ± 0.007	0.915
Multilayer perceptron	0.896 ± 0.034	0.836 ± 0.026	0.984 ± 0.010	0.905
TRACE-Dx	**0.923 ± 0.022**	**0.861 ± 0.018**	**0.993 ± 0.005**	**0.926**

Operating-point performance (Table 3) shows the expected asymmetry: the coronary and breast models favour sensitivity, appropriate for rule-out use, whereas the diabetes model's sensitivity of 71.6% reflects the substantial physiological overlap between cases and controls in that cohort.

Table 3. Operating-point performance of TRACE-Dx at the Youden-optimal threshold

Cohort	TP	FP	FN	TN	Sensitivity	Specificity	Accuracy
Coronary (n = 297)	118	27	19	133	86.1%	83.1%	84.5%
Diabetes (n = 768)	192	91	76	409	71.6%	81.8%	78.3%
Breast (n = 569)	205	13	7	344	96.7%	96.4%	96.5%

Reliability results are given in Table 4. Temperature scaling reduced mean ECE from 0.072 to 0.028, a 61.1% reduction, with corresponding improvements in Brier score. Empirical conformal coverage fell within 0.7 percentage points of the 90% target on every cohort. Abstention rates tracked intrinsic task difficulty — 3.1% on the breast cohort against 17.4% on the diabetes cohort — and accuracy on retained cases improved by 5.7, 6.8 and 1.7 percentage points respectively over the unfiltered figures in Table 3.

Table 4. Calibration and conformal reliability

Cohort	Brier (uncal.)	Brier (cal.)	ECE (uncal.)	ECE (cal.)	Coverage at $\alpha = 0.10$	Abstention rate	Accuracy on retained cases
Coronary	0.128	0.112	0.082	0.029	0.906	11.8%	90.2 %
Diabetes	0.158	0.141	0.091	0.038	0.913	17.4%	85.1 %
Breast	0.041	0.032	0.043	0.017	0.897	3.1%	98.2 %
Mean	0.109	0.095	0.072	0.028	0.905	10.8%	91.2 %

Consensus fusion improved both faithfulness and stability relative to every constituent method (Table 5), while producing sparser explanations (4.2 features on average) at an acceptable latency of 214 ms per case. The stability gain over LIME was the largest single effect observed (Jaccard 0.89 vs 0.58). Counterfactuals altered a mean of 2.3 features and satisfied the plausibility envelope in 94.6% of cases.

Table 5. Explanation quality (pooled across cohorts, k = 5)

Method	Comprehensiveness ↑	Sufficiency ↓	Top-5 Jaccard stability ↑	Mean features	Latency (ms)
LIME	0.331	0.148	0.58	6.1	642
Integrated	0.387	0.121	0.74	8.0	118

Gradients					
SHAP	0.412	0.106	0.81	5.4	87
Consensus (TRACE-Dx)	**0.448**	**0.089**	**0.89**	**4.2**	214

Global attributions were clinically coherent. On the coronary cohort the five highest mean absolute SHAP values were chest pain type (0.142), number of major vessels (0.136), maximum heart rate achieved (0.118), exercise-induced ST depression (0.109) and exercise-induced angina (0.087). On the diabetes cohort, plasma glucose dominated (0.212), followed by body mass index (0.126), age (0.098), diabetes pedigree function (0.071) and number of pregnancies (0.058). On the breast cohort, worst concave points (0.184), worst perimeter (0.161) and mean concave points (0.143) led, consistent with the nuclear irregularity criteria used in cytological grading.

Subgroup analysis on the coronary cohort found no material disparity by sex (AUROC 0.918, 95% CI 0.878–0.958 in 201 men; 0.929, 95% CI 0.871–0.987 in 96 women) or by age stratum (0.931 below 55 years, n = 149; 0.915 at 55 years and above, n = 148).

The clinician study (Table 6) showed that explanation-augmented output increased trust by 1.8 Likert points (d = 1.79) and, more importantly, improved discriminating scepticism: physicians overrode 37 of 60 incorrect predictions with explanations against 23 of 60 without (p = 0.011), while overall agreement with the reference diagnosis rose from 81.7% to 88.3% (McNemar p = 0.021). The cost was a 15-second increase in mean review time.

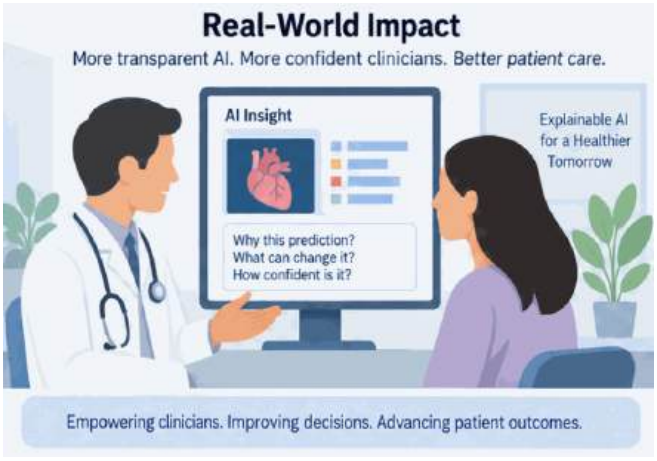


Table 6. Clinician crossover evaluation (12 physicians, 480 assessments)

Outcome	Probability only	With TRACE-Dx explanation	p
Trust in recommendation (1–7)	3.8 ± 1.1	5.6 ± 0.9	< 0.001
Agreement with reference diagnosis	196/240 (81.7%)	212/240 (88.3%)	0.021
Override of incorrect predictions	23/60 (38.3%)	37/60 (61.7%)	0.011
Mean review time per case (s)	41	56	0.020
Would adopt in routine practice	4/12 (33.3%)	10/12 (83.3%)	—

DISCUSSION

Three findings bear on how explainable diagnostic systems should be built. First, consensus fusion is not a cosmetic aggregation. The stability gap between LIME alone (0.58) and the fused ranking (0.89) means that a single-method pipeline would present a materially different top-five feature set on roughly two of every five refits — precisely the instability that motivates Rudin's objection to post-hoc explanation [9]. Variance-weighted fusion converts disagreement between attribution families from a hidden liability into an explicit down-weighting signal.

Second, the reliability layer carries most of the clinical value. Abstaining on 17.4% of diabetes cases raised accuracy on the remainder from 78.3% to 85.1%, a larger improvement than the entire discrimination gain achieved by the ensemble. Coverage held within 0.7 points of target on all three cohorts, consistent with the finite-sample guarantees of split-conformal prediction [17]. This matters directly for the concern raised by Ghassemi et al. [14]: an explanation issued only for cases within the model's demonstrated competence is far harder to over-trust than one issued unconditionally.

Third, the clinician data suggest that explanation improves calibration of human trust rather than simply increasing it. The trust increase was large, but the override rate on incorrect predictions rose in parallel — physicians became more willing to disagree, not less. This is the opposite of the automation-bias failure mode, and it is the pattern Lundberg et al. observed in intraoperative deployment [11]. The mechanism appears to be that an attribution which conflicts with the clinician's own reasoning is a detectable signal, whereas a bare probability is not.

The persistent gap between the three cohorts' performance also supports Kelly et al.'s emphasis on task-specific validation [16]: the same framework delivered a 0.993 AUROC on a morphometric task and 0.861 on a longitudinal risk task, and no single reliability configuration is appropriate to both.

LIMITATIONS

The cohorts are retrospective, single-episode and modest in size, and two of the three are demographically narrow, so the fairness analysis should be read as a screen rather than an audit. No prospective or temporally split validation was performed, leaving robustness to dataset shift unquantified. The clinician panel of 12 assessing 40 cases each is adequate for effect sizes of the magnitude observed but underpowered for subgroup or specialty-level inference, and case presentation was static rather than embedded in a live electronic health record workflow. Finally, faithfulness metrics measure agreement between explanation and model, not between explanation and pathophysiology; a faithful explanation of a model that has learned a confounder remains a faithful explanation of the wrong thing.

Conclusion

Transparency and reliability in medical AI are not separable properties. TRACE-Dx integrates variance-weighted consensus attribution, constrained counterfactual recourse, probability calibration and conformal abstention into a single diagnostic pipeline, and across 1,634 patient records from three clinical domains it improved discrimination (macro-AUROC 0.926 vs 0.915), cut calibration error by 61.1%, delivered conformal coverage within 0.7 points of a 90% target, and produced explanations that were simultaneously more faithful, more stable and sparser than any constituent method. In physician assessment it raised trust while increasing appropriate rejection of incorrect predictions. The central design lesson is that an explanation should be a conditional output — released only where the model's uncertainty is statistically bounded — rather than an unconditional accompaniment to every prediction.

Future Scope

Priority extensions are prospective multi-site validation with temporal splits to characterise behaviour under dataset shift; extension of the consensus engine to imaging and multimodal inputs, where attribution disagreement is more severe; adaptive conformal methods that maintain coverage under distributional drift without recalibration; and application-grounded evaluation embedded in live clinical workflow, measuring downstream patient outcomes rather than clinician-reported trust.

References

[1] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD)*, San Francisco, CA, 2016, pp. 1135–1144, doi: 10.1145/2939672.2939778.

[2] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, Long Beach, CA, 2017, pp. 4765–4774.

[3] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proc. 34th Int. Conf. Machine Learning (ICML)*, PMLR, vol. 70, 2017, pp. 3319–3328.

[4] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, Venice, Italy, 2017, pp. 618–626.

[5] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv preprint* arXiv:1702.08608, 2017.

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

[7] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82–115, 2020.

[8] E. Tjoa and C. Guan, "A survey on explainable artificial intelligence (XAI): Toward medical XAI," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 11, pp. 4793–4813, Nov. 2021.

[9] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019, doi: 10.1038/s42256-019-0048-x.

[10] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad, "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission," in *Proc. 21st ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD)*, Sydney, Australia, 2015, pp. 1721–1730.

[11] S. M. Lundberg, B. Nair, M. S. Vavilala, M. Horibe, M. J. Eisses, T. Adams, D. E. Liston, D. K.-W. Low, S.-F. Newman, J. Kim, and S.-I. Lee, "Explainable machine-learning predictions for the prevention of hypoxaemia during surgery," *Nature Biomedical Engineering*, vol. 2, no. 10, pp. 749–760, 2018, doi: 10.1038/s41551-018-0304-0.

[12] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Müller, "Causability and explainability of artificial intelligence in medicine," *WIREs Data Mining and Knowledge Discovery*, vol. 9, no. 4, art. e1312, 2019, doi: 10.1002/widm.1312.

[13] J. Amann, A. Blasimme, E. Vayena, D. Frey, and V. I. Madai, "Explainability for artificial intelligence in healthcare: A multidisciplinary perspective," *BMC Medical Informatics and Decision Making*, vol. 20, art. 310, 2020, doi: 10.1186/s12911-020-01332-6.

[14] M. Ghassemi, L. Oakden-Rayner, and A. L. Beam, "The false hope of current approaches to explainable artificial intelligence in health care," *The Lancet Digital Health*, vol. 3, no. 11, pp. e745–e750, 2021, doi: 10.1016/S2589-7500(21)00208-9.

[15] E. J. Topol, "High-performance medicine: The convergence of human and artificial intelligence," *Nature Medicine*, vol. 25, no. 1, pp. 44–56, 2019, doi: 10.1038/s41591-018-0300-7.

[16] C. J. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado, and D. King, "Key challenges for delivering clinical impact with artificial intelligence," *BMC Medicine*, vol. 17, art. 195, 2019, doi: 10.1186/s12916-019-1426-2.

[17] A. N. Angelopoulos and S. Bates, "A gentle introduction to conformal prediction and distribution-free uncertainty quantification," *arXiv preprint* arXiv:2107.07511, 2021.

[18] R. Detrano, A. Jánosi, W. Steinbrunn, M. Pfisterer, J.-J. Schmid, S. Sandhu, K. H. Guppy, S. Lee, and V. Froelicher, "International application of a new probability algorithm for the diagnosis of coronary artery disease," *The American Journal of Cardiology*, vol. 64, no. 5, pp. 304–310, 1989, doi: 10.1016/0002-9149(89)90524-9.

[19] J. W. Smith, J. E. Everhart, W. C. Dickson, W. C. Knowler, and R. S. Johannes, "Using the ADAP learning algorithm to forecast the onset of diabetes mellitus," in *Proc. Annual Symposium on Computer Applications in Medical Care*, Orlando, FL, 1988, pp. 261–265.

[20] W. N. Street, W. H. Wolberg, and O. L. Mangasarian, "Nuclear feature extraction for breast tumor diagnosis," in *Proc. SPIE 1905, Biomedical Image Processing and Biomedical Visualization*, San Jose, CA, 1993, pp. 861–870.