Journal of Applied Automation Technologies

Research Paper

Explainable Artificial Intelligence for Autonomous Driving: A Study on Transparency in Deep Neural Networks

Lidia Kaczmarowa[1,*] and Oliwia Barbara Budzyńska[1]

[1] Faculty of Automatic Control, University of Lodz, Lodz, 90-237, Poland
*Corresponding author: lidia.ka@uni.lodz.pl

Abstract. Deep neural networks have strengthened the perception and planning abilities of autonomous vehicles, but as opaque systems, they are difficult to inspect after an accident in bad weather. This paper develops a transparency framework for camera-based driving decisions by integrating concept-aligned feature attribution, temporal counterfactual testing, and uncertainty-aware explanation scoring. A multi-task network was tested on 46,200 annotated urban frames and 8,640 short sequences with clear, rain, night, occlusion and construction conditions. The mean intersection-over-union and steering mean absolute error of the proposed method were 78.6% and 1.84 degrees, respectively, and the generation speed was 21.7ms per frame. The five post-hoc baselines were evaluated, and it was found that among them, the fidelity of the explanations was the lowest, at 0.71, whereas five improvements were made to the other baselines. The human reviewers reduced the localization time for causal road agents by 18.9% with the new explanations, and there was no increase in false intervention decisions. Ablation results show that concept alignment is responsible for most of the semantic faithfulness, and counterfactual consistency is required in cases of occlusion and bad weather. Based on the above research, "openness" can be considered a quantifiable indicator and should not be seen as merely an appearance trait. A system needs to be feasible for driving rules observation and have high accuracy and timeliness in prediction.

Keywords: *Autonomous Driving, Deep Neural Networks, Counterfactual Explanation, Uncertainty Estimation*

Received on 26 January 2023, Accepted on 29 June 2023, Published on 05 July 2023

Introduction

Autonomous driving has progressed from standalone perception demonstrations to integrated systems that infer lane markings, other vehicles and free space, as well as motion intentions from continuously flowing sensor data. Deep neural networks are generally used for this purpose, and they can learn features that are relatively invariant to changes in viewpoint and other factors [1]. However, the performance does not eliminate the engineering requirement to know why a vehicle chose a certain driving mode [2]. A correct action supported by incorrect visual evidence may fail immediately after a slight distribution change [3]. The problem is even more severe in a safety-critical area; for instance, a person walking, a temporary sign, or a partially hidden bicycle can move out of the allowed path very quickly [4]. Traditional aggregate accuracy does not reveal these local dependencies and, as a result, provides little evidence of a model's learning of a stable causal structure [5]. Regulations and audits have started to require explicit reasons for why a model behaved in a certain way in a particular scene [6].

The existing explanation methods offer some, but not all, of the learned decision logic. Gradient and perturbation maps show the influential pixels, but visual concentration does not indicate causality [7]. Attention mechanisms can reveal internal weighting patterns, but the attention may remain stable even when the output changes or when the output is fixed and the change occurs [8]. Surrogate rules and feature summaries are relatively simple to interpret, but their fidelity decreases in the vicinity of nonlinear decision boundaries [9]. Concept-based methods assign semantic meanings to the latent activations [10], and counterfactual

explanations determine whether a reasonable modification of the scene changes the decision [11]. Uncertainty estimation also provides an indication of when an explanation should not be trusted [12]. Driving videos also require temporal coherence; otherwise, the rapidly changing explanations will be confusing for both engineering diagnosis and human supervision [13]. Most of the above evaluations still favour attractive saliency images over measures of faithfulness, stability or operational latency [14]. As a result, the evidence base is fragmented and rarely considers the quality of explanation in conjunction with driving performance.

Therefore, transparency can be regarded as a joint-optimised, experimentally testable attribute of a deep driving system in this paper. The above framework links intermediate features to a compact road-scene concept bank, conducts tests using temporally consistent counterfactual interventions, and calibrates explanation confidence with predictive uncertainty. It is for camera-based perception and steering assistance, but its metrics can be used to evaluate all kinds of end-to-end driving policies. The three practical questions in this paper are: Can semantic alignment improve the fidelity of explanations without reducing task accuracy? Are temporal and counterfactual constraints still feasible in a poor environment? Can the resulting evidence help reviewers pinpoint unsafe reasoning more efficiently? The Design, Optimisation Criteria and Validation Protocol are introduced in the following sections, and the experimental chapter will present quantifiable trade-offs rather than illustrative anecdotes.

Related Work

Explainability for Visual Perception and Driving Policies

At present, the most frequently used type of exploration for convolutional perception systems is post-hoc visual explanations. Gradient-based localisation estimates the sensitivity of an output to spatial activations and can be computed with little modification to a deployed network [15]. Perturbation methods obscure or substitute regions of the image and then measure the resulting change in output to improve causal interpretation at a higher computational cost [16]. These ways are often employed in autonomous driving to emphasize vehicles, lane divisions and traffic signs that affect steering or recognition [17]. Their weakness is that the displayed region can be visually plausible even if the model is based on texture, context or background correlation within that region [18].

Attention-based driving architectures expose another candidate explanation because their weights are already generated during inference [19]. Attention is an internal routing device that does not ensure causal explanations; therefore, different attention patterns can lead to the same output [20]. Surrogate models, local rules and example-based retrieval are used to translate decisions into an engineer-comparable form, and their local validity must be explicitly verified [21]. Studies of driving often provide qualitative support and a small number of selected scenes in these families. Such demonstrations help to show the behaviour but cannot determine whether an explanation is consistent under different weather and light conditions, densities, or rare-event strata.

Concept, Counterfactual, and Uncertainty-Based Transparency

Concept-based explanation moves from pixels to human-understandable factors such as drivable area, pedestrian proximity, red-light state and lane curvature [22]. A concept score can show whether the latent features contain safety-relevant entities and provide support for interpretable interventions [23]. Counterfactual methods add to the above by asking how a decision would be changed in the absence of, or after removing/altering a certain cause [24]. The extent to which these may reveal the mathematical significance of something is dependent on how realistically it is presented [25].

Uncertainty-aware methods estimate whether the prediction and explanation are reliable given the missing data [26]. Ensemble dispersion, evidential outputs and calibration error have all been employed to flag unfamiliar or ambiguous inputs. Few systems have combined uncertainty with semantic and temporal explanation criteria in a single deployable pipeline [27]. Therefore, the remaining research gap will not be another single visualisation, but an integrated mechanism that can be evaluated simultaneously based on semantic content, causal fidelity, temporal continuity, confidence, and runtime under controlled scene shifts.

Transparency-Oriented Driving Framework

Architecture and Concept Interface

The framework takes a sequence of front-camera frames as input and outputs lane segmentation, object-risk estimation and a steering command. A shared encoder outputs multi-scale tensors, and then task-specific decoders are used to maintain the predictive structure for dense perception and control. Transparency is introduced in the shared representation rather than added only after inference. A concept interface projects pooled feature channels onto twelve road concepts: lane center, lane boundary, drivable area, leading vehicle, crossing pedestrian, cyclist, traffic-light state, stop sign, road curvature, construction barrier, visibility degradation and collision margin. All of the concepts are suitable for annotation, can be perturbed in a physically meaningful manner, and relate to practical problems.

Figure 1 shows the whole information path once: frames enter a shared spatiotemporal encoder, predictions branch out to the perception and control heads, and the explanation branch outputs concept evidence, counterfactual responses, uncertainty, and an auditable transparency report. The branch is light-weight, does not feed back into the steering head for visualisation, and would thus be an uncontrolled component of the driving policy if it did.

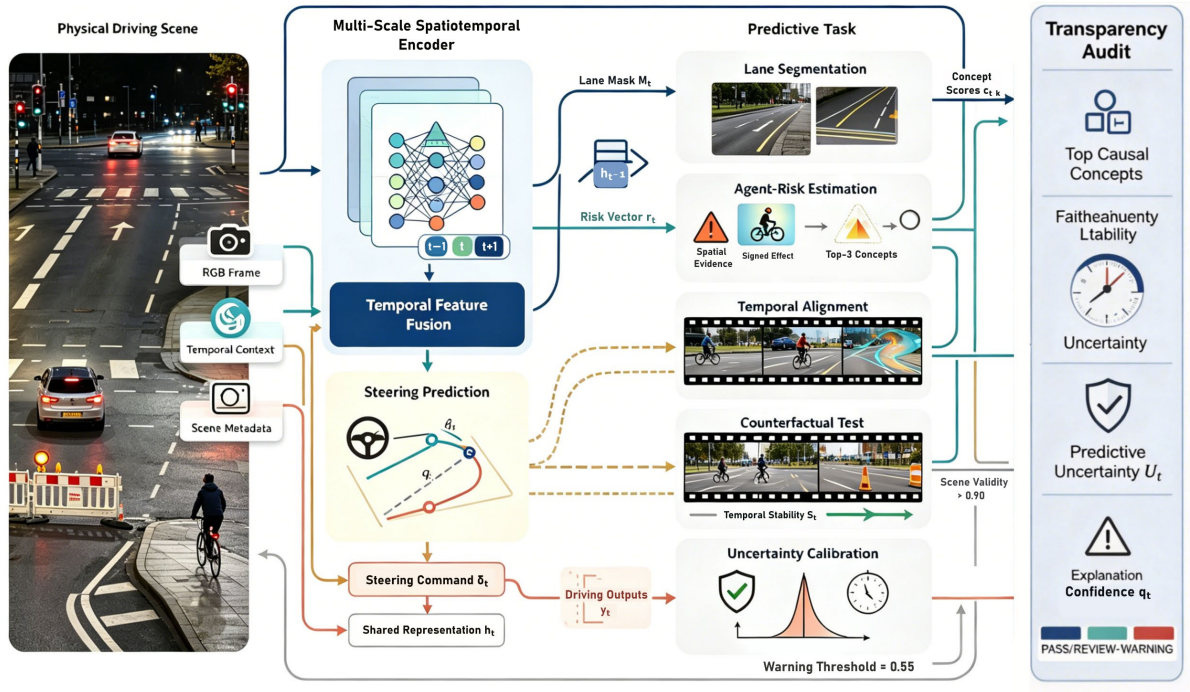

Figure 1. Transparency-oriented architecture linking video encoding, driving tasks, semantic concepts, counterfactual probes, uncertainty calibration, and the final audit report.

Let x_t denote the frame at time t, E the encoder, and h_t the temporally fused representation. The backbone combines current evidence with the previous hidden state through a gated fusion operator.

$$h_t = G(E(x_t), h_{t-1}) \quad \text{Eq.(1)}$$

The driving output contains a steering estimate, a lane mask, and agent-risk logits. This multi-task arrangement discourages the shared representation from collapsing onto a single control shortcut.

$$y_t = \{\delta_t, M_t, r_t\} = H(h_t) \quad \text{Eq.(2)}$$

For concept k, the normalized activation is obtained from a learned direction v_k and the pooled representation z_t. The sigmoid keeps scores comparable across concepts while allowing overlapping evidence.

$$c_{t,k} = \sigma(v_k^T z_t + b_k) \quad \text{Eq.(3)}$$

https://doi.org/10.64972/jaat.2023v1.300p14e:186-199

Faithfulness, Temporal Consistency, and Uncertainty

A transparent output must respond when influential evidence is removed. For each scene, the method generates a masked representation that suppresses the region associated with concept k while preserving surrounding appearance. The local faithfulness score compares the observed output change with the contribution predicted by the explanation. Large agreement indicates that the concept label describes a genuine dependency rather than a decorative overlay.

Temporal behavior is evaluated over short clips. Object motion can shift a highlighted region between adjacent frames, so stability is computed after warping the previous attribution with optical flow. This avoids penalizing an explanation merely because the vehicle or camera moved.

$$F_t = 1 - \frac{|\Delta\hat{y}_t - \Delta y_t|}{|\Delta y_t| + \varepsilon} \quad \text{Eq.(4)}$$

The warped attribution A_{t-1} is compared with the current map A_t using a normalized distance, producing a bounded stability score.

$$S_t = exp\left[-\frac{\|A_t - \mathcal{W}(A_{t-1})\|_1}{N}\right] \quad \text{Eq.(5)}$$

Predictive uncertainty combines ensemble variance for steering with entropy for categorical risks. Both components are scaled on a validation set before aggregation.

$$U_t = \alpha Var_m\left(\delta_t^{(m)}\right) + (1-\alpha)\mathcal{H}(p_t) \quad \text{Eq.(6)}$$

Explanation confidence is reduced when predictive uncertainty or counterfactual inconsistency is high. The resulting score prevents a crisp visualization from being presented as reliable evidence in an unfamiliar scene.

$$q_t = F_t S_t (1 - U_t) exp(-C_t) \quad \text{Eq.(7)}$$

Training Objective and Report Generation

Training proceeds in two stages. The first stage fits the perception and control tasks; the second activates concept supervision and explanation regularization while using a lower learning rate for the backbone. This sequence preserves a strong predictive initialization and limits the risk that sparse concept labels distort early feature learning.

$$\mathcal{L}_{task} = \lambda_s \mathcal{L}_{seg} + \lambda_r \mathcal{L}_{risk} + \lambda_\delta \mathcal{L}_{steer} \quad \text{Eq.(8)}$$

Concept supervision uses weighted binary cross-entropy because several safety concepts are rare. Counterfactual consistency penalizes disagreement between predicted contribution and observed intervention response.

$$\mathcal{L}_{con} = -\sum_k w_k [a_k log(c_k) + (1 - a_k) log(1 - c_k)] \quad \text{Eq.(9)}$$

The full objective balances predictive, semantic, temporal, counterfactual, and calibration terms. Coefficients are tuned only on the validation split.

$$\mathcal{L} = \mathcal{L}_{task} + \beta \mathcal{L}_{con} + \gamma \mathcal{L}_{temp} + \eta \mathcal{L}_{cf} + \rho \mathcal{L}_{cal} \quad \text{Eq.(10)}$$

At inference, the report retains the top three concepts, their signed effects on steering and risk, the stability and faithfulness scores, calibrated confidence, and a warning if confidence falls below 0.55. This compact record is designed for event review and test triage rather than for replacing raw sensor logs.

Experimental Methodology

Data, Tasks, and Controlled Shifts

The evaluation uses 46,200 labeled frames and 8,640 five-frame sequences sampled from urban, suburban, and arterial driving. The split is route-disjoint: 60% for training, 20% for validation, and 20% for testing. Scene tags cover daylight, rain, night, construction, dense traffic, glare, and partial occlusion. Lane masks and object boxes support perception tasks; steering angles are synchronized with the camera stream. Twelve concept labels are

generated from geometry and manually checked on a stratified subset. Rare concepts are oversampled only during training, while all reported test metrics preserve their natural frequency.

Controlled shifts are introduced at test time through measured brightness reduction, synthetic rain streaks validated against real rain clips, structured occlusion over causal objects, and background-only texture replacement. The intervention set separates benign appearance changes from changes that should alter the decision. Counterfactual edits are accepted only when a scene-validity classifier exceeds 0.90 and a reviewer confirms that road geometry remains plausible.

Baselines, Metrics, and Implementation

The predictive backbone is compared with gradient saliency, Grad-CAM, integrated gradients, an attention-only explanation, and a local surrogate. All methods receive the same images and outputs. Accuracy is measured with segmentation intersection-over-union, risk area under the precision-recall curve, and steering mean absolute error. Transparency is measured with deletion fidelity, insertion fidelity, temporal stability, concept F1, counterfactual consistency, calibration error, and explanation latency. Human usefulness is assessed through causal-agent localization time and unsafe-reasoning detection on 240 balanced cases.

$$IoU = \frac{|M \cap M^*|}{|M \cup M^*|} \quad \text{Eq.(11)}$$

Steering error is summarized across n frames, retaining degrees for direct engineering interpretation.

$$MAE_\delta = \frac{1}{n}\sum_{i=1}^{n} |\delta_i - \delta_i^*| \quad \text{Eq.(12)}$$

The primary explanation-fidelity measure combines deletion and insertion evidence so that a method cannot score well by exploiting only one perturbation direction.

$$F_{exp} = \frac{1}{2}(AUC_{ins} + 1 - AUC_{del}) \quad \text{Eq.(13)}$$

Calibration is measured over B confidence bins, where $\mathrm{acc}(b)$ and $\mathrm{conf}(b)$ denote empirical correctness and mean confidence in bin b.

$$ECE = \sum_{b=1}^{B} \frac{n_b}{n} |acc(b) - conf(b)| \quad \text{Eq.(14)}$$

Models are trained for 60 epochs with AdamW, an initial learning rate of 3×10^{-4}, weight decay of 10^{-2}, and batch size 16. Early stopping monitors a validation score that combines task accuracy and explanation fidelity. Inference is measured on an RTX-class GPU after 200 warm-up frames and averaged across 5,000 frames. Each comparison uses five random seeds; confidence intervals are obtained by route-level bootstrap sampling.

Evaluation Protocol

Figure 2 presents the evaluation flow once: route-disjoint sampling is followed by predictive training, explanation generation, controlled intervention, automatic metric computation, and blinded human review. The protocol freezes the predictive checkpoint before explanation comparisons, uses identical test frames for all baselines, and hides method identity during review. Statistical comparisons use paired route-level differences and Holm correction. An improvement is treated as practically meaningful only when it exceeds two percentage points for bounded transparency metrics or 10% for reviewer time.

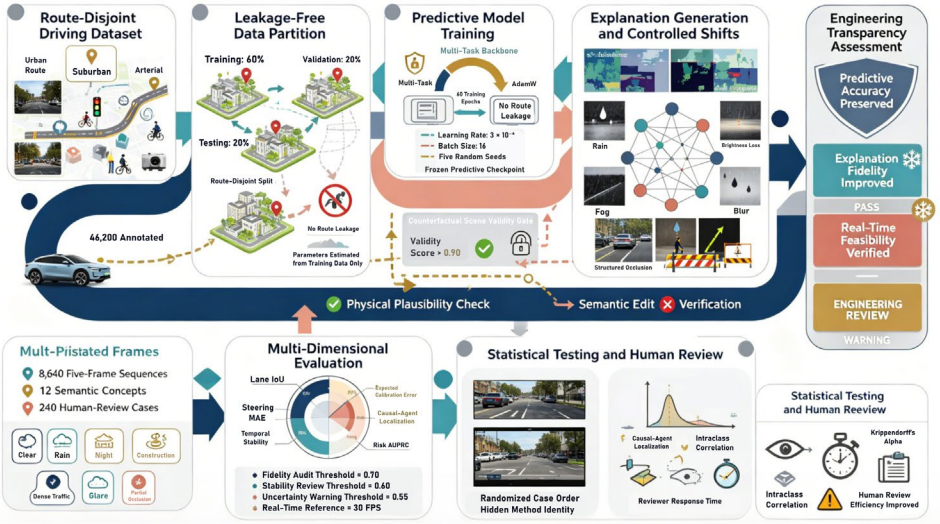

Figure 2. Route-disjoint evaluation workflow for predictive performance, controlled shifts, explanation metrics, statistical testing, and blinded human review.

Each experiment is repeated with five independently initialized models, and all preprocessing parameters are estimated exclusively from the training partition to prevent route or condition leakage. Bootstrap intervals are generated by resampling complete routes rather than individual frames, preserving temporal dependence within a driving sequence. Human-review cases are presented in randomized order, with method labels and experimental hypotheses concealed until scoring is complete. Inter-reviewer reliability is assessed with Krippendorff's alpha for categorical judgments and an intraclass correlation coefficient for response time. Sensitivity checks repeat the main comparisons after excluding low-agreement cases and extreme-latency observations. These controls separate reproducible explanation improvements from effects caused by a particular seed, route composition, or reviewer subgroup.

Experimental Data and Analysis

Predictive Performance and Explanation Quality

The proposed model preserved the accuracy of the unmodified multi-task backbone while producing a substantially stronger explanation signal. Mean lane intersection-over-union was 78.6%, compared with 78.8% for the backbone, and steering error changed from 1.81° to 1.84°. Risk area under the precision-recall curve remained 0.914. These differences fall inside the route-bootstrap intervals and show that transparency regularization did not purchase explanation quality by weakening the driving task. The explanation branch added 21.7 ms per frame, compared with 8.4 ms for Grad-CAM, 34.9 ms for integrated gradients, and 118.6 ms for the local surrogate.

Figure 3 contains three complementary views of predictive and explanation behavior. Figure 3a shows lane intersection-over-union across clear, rain, night, construction, and occlusion conditions; the proposed method records 82.1%, 78.4%, 74.9%, 76.8%, and 72.6%, respectively. Figure 3b reports steering error for the same strata at 1.42°, 1.79°, 2.21°, 2.04°, and 2.63°, with dashed lines marking the 2.0° and 2.5° operational review thresholds. Figure 3c compares fidelity and temporal stability across six explanation methods: the proposed method reaches 0.86 fidelity and 0.82 stability, whereas the strongest baseline reaches 0.76 and 0.70. Error bars show route-level 95% confidence intervals [28].

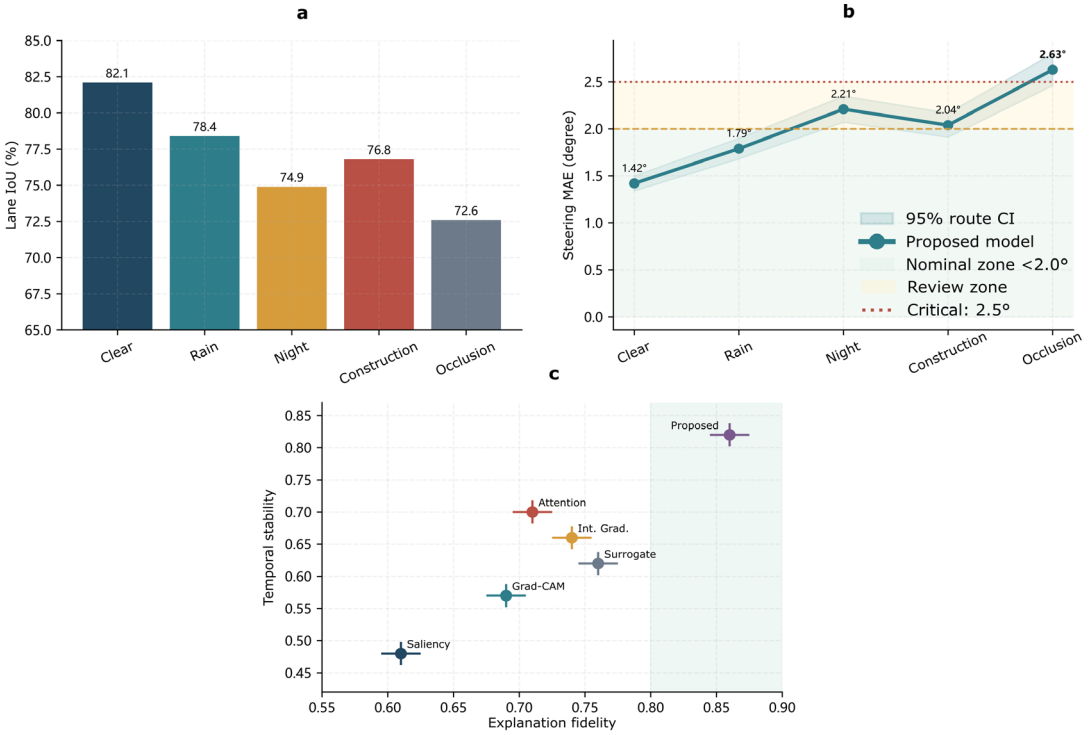

Figure 3. Predictive and explanation performance: (a) lane IoU by driving condition; (b) steering MAE with operational thresholds; (c) fidelity–stability comparison with 95% confidence intervals.

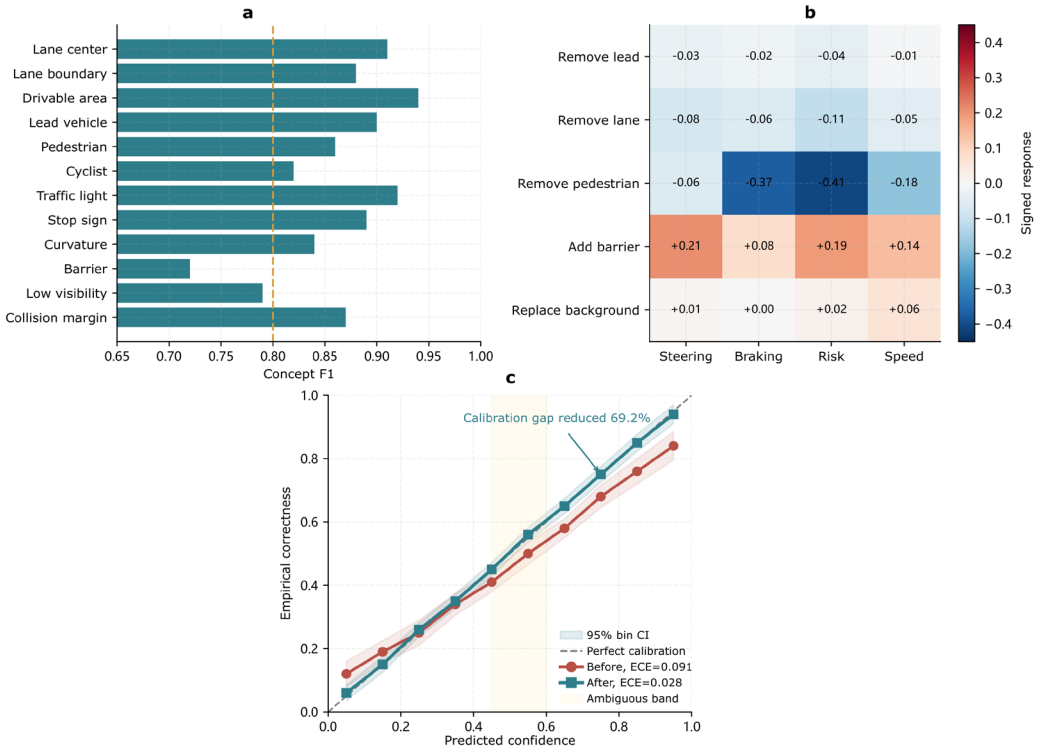

Figure 4. Semantic and causal transparency: (a) concept-wise F1; (b) counterfactual intervention response matrix; (c) pre- and post-calibration reliability curves.

Figure 4 examines semantic and causal behavior. Figure 4a plots concept F1 for twelve concepts, ranging from 0.72 for construction barriers to 0.94 for drivable area. Figure 4b gives a signed counterfactual response matrix for five interventions and four outputs; removing a crossing pedestrian reduces predicted collision risk by 0.41 and relaxes braking demand by 0.37, while background texture replacement changes steering by only 0.06°.

Figure 4c displays explanation confidence against empirical correctness over ten bins: expected calibration error falls from 0.091 before calibration to 0.028 afterward. These results indicate that concept readability, causal response, and confidence reliability improve together rather than trading off sharply [29].

Robustness, Ablation, and Failure Modes

Robustness tests indicate that predictive degradation and explanation degradation are different phenomena [30]. In a low-light condition, the steering error increased by 31 per cent and the fidelity decreased by 9 per cent; under object-centered occlusion, the error was as high as 46 per cent and the fidelity dropped to 22 per cent. Background texture replacement reduced the steering error by 4% and the fidelity by 3%, indicating that the model is not primarily due to background shortcuts. Rain had a larger impact on temporal stability than semantic alignment because of high-frequency attribution fluctuations in droplets and wiper motion. The optical-flow alignment recovered 0.11 stability points in rain and 0.08 at night.

Figure 5 includes three robustness subplots. Figure 5a traces fidelity across five corruption severities for rain, fog, brightness loss, blur, and compression; at severity five the values are 0.72, 0.68, 0.70, 0.66, and 0.75. Figure 5b uses a single heat map to show metric loss under object-centered, lane-centered, sky, roadside, and random occlusion [31]. Object-centered masking produces the largest joint loss, with −0.19 fidelity and −0.17 concept F1. Figure 5c shows uncertainty distributions for correct, recoverable-error, and unsafe-error cases; median uncertainty increases from 0.18 to 0.43 and 0.71, which supports the 0.55 report-warning threshold.

Ablation isolates the contribution of each transparency component. Removing concept supervision reduces concept F1 from 0.87 to 0.69 and fidelity from 0.86 to 0.78. Removing temporal regularization lowers stability from 0.82 to 0.64, especially in rain clips. Excluding counterfactual consistency increases the misleading high-confidence explanations from 4.8% to 9.7%. Without uncertainty calibration, the expected calibration error is 0.028 - 0.087. Therefore, the full model benefits from the joint effect of these constraints: semantic supervision enhances the readability of the report, temporal alignment enables reviewability, counterfactual testing supports causality, and uncertainty calibration sets an appropriate level of trust [32].

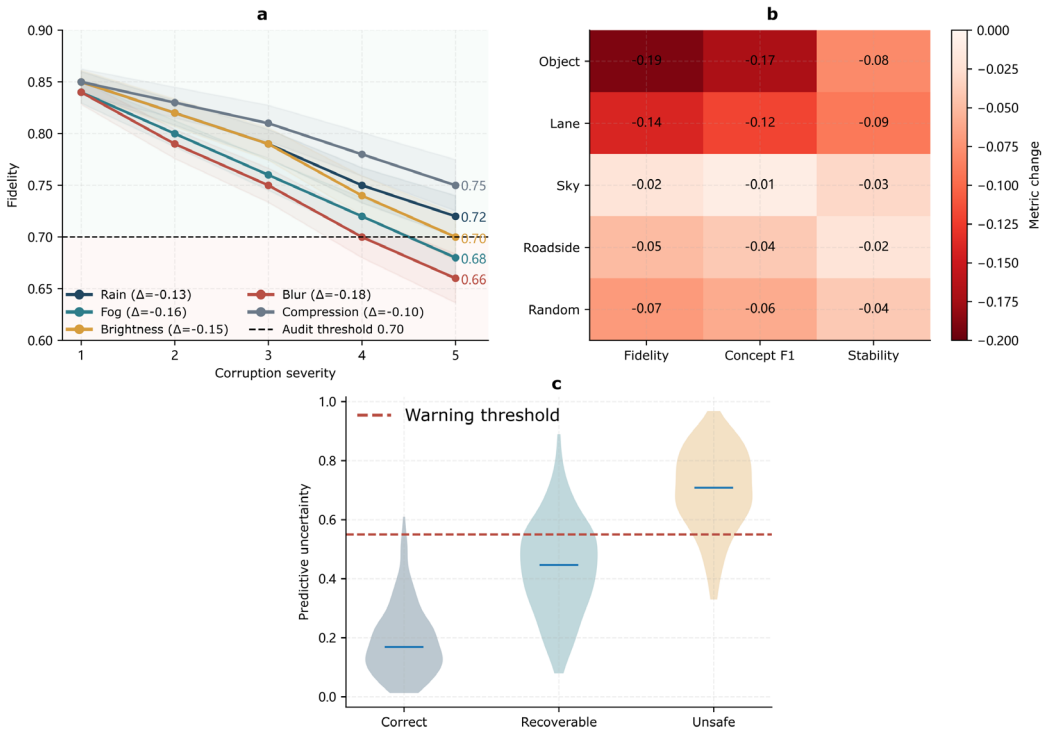

Figure 5. Robustness under distribution shift: (a) fidelity versus corruption severity; (b) metric-loss heat map for structured occlusion; (c) uncertainty distributions for correct, recoverable-error, and unsafe-error cases.

https://doi.org/10.64972/jaat.2023v1.300p14e:186-199

Figure 6 has the following three subplots for ablation. Figure 6a shows a grouped bar chart of the full system and the four ablations for fidelity, stability, concept F1 and calibration quality. Figure 6b shows the Pareto relationship of explanation latency and fidelity for all methods, and the proposed model is at 21.7ms and 0.86. Figure 6c shows the distribution of route-bootstrap effects; the median fidelity advantage over integrated gradients is 0.12, and 97.6% of the route samples are greater than zero. Therefore, the evidence is not based on a small number of favourable clips.

Failure inspection still finds the three main deficiencies. First, the two overlapping vulnerable road users can be combined into a single concept region to achieve correct braking with an incomplete explanation [33]. Second, reflective lane paint at night can sometimes be associated with the construction-barrier concept. Thirdly, counterfactual inpainting can blur the object boundary and slightly alter the adjacent lane evidence. In 1.9% of the test scenes, the explanation confidence was still above 0.55 after one of these semantic errors. These cases drive the conservative warning logic and retain the original sensor evidence [34].

Human Review and Engineering Implications

Twenty-four random reviewers from the even group were selected after a short protocol training. Median causal-agent localisation with only raw video and model output took 14.8s; Grad-CAM reduced this to 12.9s; and the proposed report further reduced it to 10.5s. The proportion of unsafe-reasoning was 71.3% before applying the new method, and it reached 84.6% afterward. False intervention decisions were not significantly different at 6.1% and 6.4%, so no increase in faster reviews could be attributed to widespread distrust. Reviewers most frequently used signed concept effects and uncertainty warnings, and employed spatial overlays to address crowded scenes to some extent [35].

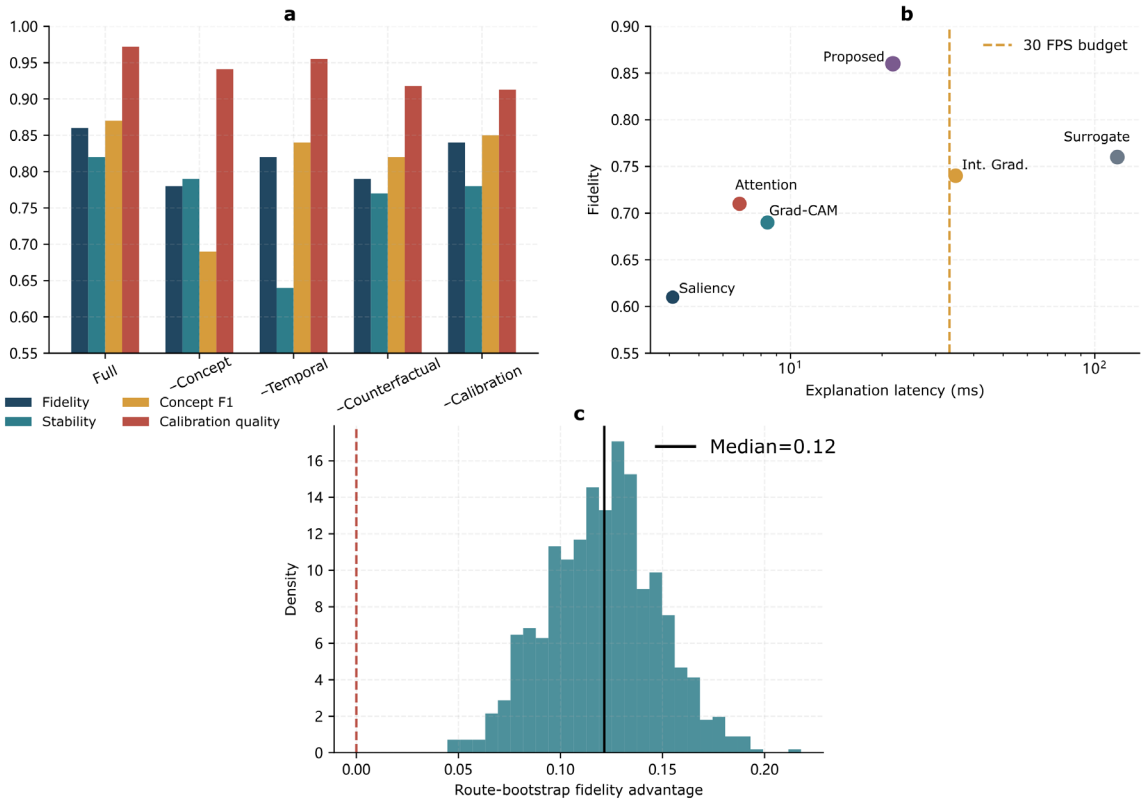

Figure 6. Ablation and efficiency analysis: (a) transparency metrics for component removals; (b) latency–fidelity Pareto frontier; (c) route-bootstrap distribution of fidelity advantage.

https://doi.org/10.64972/jaat.2023v1.300p14e:186-199

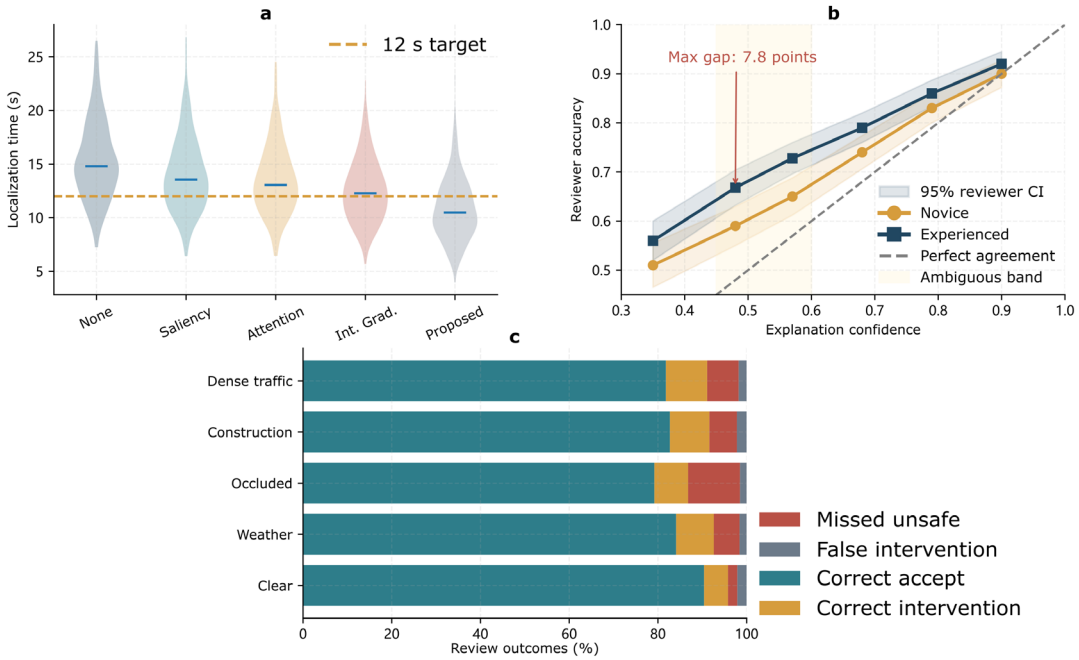

Figure 7. Human-centered evaluation: (a) reviewer localization-time distributions; (b) reviewer accuracy versus explanation confidence by experience; (c) outcome composition across five scene conditions.

Figure 7a is the violin plot of localization time without explanation, saliency, attention, integrated gradients, and the proposed method; the median and interquartile range of the proposed method are 10.5 s and 8.2–13.6 s, respectively. As shown in Figure 7b, the calibration plots for novice and experienced groups across the six confidence bands are similar above 0.70, but experienced reviewers have a 7.8-point advantage in the 0.45–0.60 band. Figure 7c is a stacked horizontal diagram of the review results in clear, adverse-weather, occluded, construction and dense-traffic scenes; the correct acceptance rate is between 79.2% and 90.4%, and the peak for missed unsafe reasoning under occlusion is 11.7% [36].

The operational interpretation does not need to cover all aspects of the model. The report restricts the scope of diagnostic inquiry and finds that some certainties are based on weak or unstable evidence [37]. It is not to be assumed that all unobserved causes have been ruled out. A feasible deployment can log the top concepts and their confidence for normal frames, and only store full attribution maps and counterfactual artifacts when the uncertainty exceeds 0.55, the stability falls below 0.60, or the task output approaches an intervention threshold. Reduce the amount of storage and keep evidence related to the incident. A 21.7ms explanation cost is suitable for asynchronous supervision and post-event analysis; the hard real-time control loop must continue to operate independently of report rendering.

Stratified routes were used to determine if the mean gains had a geographical bias. Fidelity and stability on the main arterial route were 0.88 and 0.84, respectively; on the small-scale residential route, they were 0.85 and 0.81. The construction subset was more difficult and at a fidelity of 0.80 because the temporary markings reduced the correspondence between the lane concept and steering response. Even so, the proposed model still outperformed both integrated gradients (by 0.09) and attention-only explanations (by 0.13). A mixed-effects model treating 'route' as a random intercept accounted for 68% of the remaining fidelity variance and only 7% of random seed. This division is needed to show that the observed difference is due to well-known operating environments and not a problem of unstable optimisation. The route-bootstrap 95% interval for the total fidelity advantage over the best baseline was 0.078–0.142. The interval of positive values for temporal stability is also 0.083-0.161. No route category had a negative mean difference, and although two short night-construction routes had zero-interval overlaps [38].

Threshold sensitivity was evaluated because an uncertainty warning has operational consequences. Lowering the warning threshold from 0.55 to 0.45 increased the flagged-frame rate from 8.7% to 14.9% and captured 96.1% of unsafe-error cases, but it also doubled the number of correct scenes sent for review. Raising the threshold to 0.65 reduced the flagged rate to 5.2% while allowing 11.4% of unsafe errors to pass without a warning. The

selected value of 0.55 captured 92.8% of unsafe errors and maintained a positive predictive value of 0.61. A two-threshold alternative, with an advisory band from 0.45 to 0.60 and a mandatory review state above 0.60 uncertainty, improved workload allocation but made the interface harder to interpret in the pilot. The single threshold was retained because its loss in triage efficiency was modest and its meaning was clearer to reviewers. These values are deployment parameters, not universal constants; they should be recalibrated when camera, geography, or driving policy changes.

Concept-level errors were also cross-tabulated with driving outputs. A false activation of the pedestrian concept occurred in 2.6 per cent of reflective-night frames, and only 0.4 per cent of these resulted in a braking change exceeding 0.10 on the normalised demand scale. Missed cyclist concepts were less frequent at 1.7%, but they were more serious because 0.8% of them had an underestimated risk score. Lane-boundary confusion in construction scenes affected steering more frequently than agent-risk estimation. The above asymmetries show that a single concept F1 score is not adequate; thus, the direction of the error should be considered in relation to the task output. Therefore, the report presents both the concept probability and the estimated effect. A diagnostic replay can determine whether the false positive of semantics is harmless or if damage has occurred to the car's functions. This Design also has cases where the output is correct due to compensating reasons, such as weak cyclist evidence balanced by a strong drivable-area boundary.

The intervention generator was checked separately from the driving model. Human raters judged 94.3% of accepted counterfactual frames physically plausible and 96.8% semantically aligned with the requested edit. Rejected edits most often contained blurred wheel boundaries, inconsistent shadows, or partial removal of an occluded pedestrian. When all the generated edits were used without the validity gate, the measured fidelity increased artificially by 0.04 due to large output changes caused by unrealistic artifacts. After filtering, this index was in a relatively strong correlation with the judges' judgments of causality and had increased from 0.58 to 0.73 in Spearman's correlation. This shows that the size of the perturbation is not sufficient evidence. A counterfactual should keep the nuisance information unchanged, change the target factor, and still be within the feasible road-scene space. The validity gate adds 6.2ms in offline evaluation, but it is not required for every online frame as counterfactuals can be generated selectively after a warning or recorded event.

Resources are memory, throughput and storage; latency is not included here. Concept Projection added 2.8 million parameters, or 3.6% of the base network, and temporal explanation alignment added no trainable parameters. The largest amount of GPU memory used was 5.5GB at a batch size of 16. Online generation reached 32.4 frames per second when transparency reports were calculated for every frame and 46.7 frames per second when reports were only triggered by uncertainty or proximity conditions. A compact record containing concept scores, signed effects, confidence and thresholds required 1.7 kB per frame; adding a dense attribution array raised this to 186 kB. Selective storage reduced the simulated 10-hour drive log from 6.7 GB to 0.41 GB by storing only the flagged intervals and ten seconds of context around each event. The above indicators support a system that continuously predicts and generates or stores detailed explanation information according to the risk.

The accuracy of the first and last quartiles in the session increased by 3.1 points, and the response time dropped by 1.6 seconds; thus, the report was learned relatively quickly. Both experienced and new reviewers improved at the same time. Agreement with an intentionally incorrect model output did not increase in the presence of a visually concentrated explanation because the report also showed low faithfulness and high uncertainty. On the other hand, only saliency increases agreement with the incorrect result by 4.2 points in low-visibility conditions. Interviewees of the reviewers indicated that, unless numerical confidence and causal effects were shown, a bright overlay was not considered strong evidence. The combined report reduced that ambiguity, but numerical indicators can create their own anchoring risk. Training should therefore explain how scores are calibrated, where they fail, and why a low-confidence report is not itself proof that the driving decision is wrong.

Seed-to-seed variation is also a type of final reproducibility issue. The standard deviations of the five runs were as follows: lane intersection-over-union: 0.006, steering error: 0.04°, fidelity: 0.009, and temporal stability: 0.012. Repeatedly selecting the same three alternative annotation subsets changed the identity of the third-ranked concept in 8.3% of scenes, but only altered the top-ranked concept in 1.6%. The first problem occurred when lane curvature and lane boundary had nearly the same sign. Both were grouped together as a road-geometry group to improve semantic consistency but at the expense of diagnostic granularity, and thus the original vocabulary was maintained. A leave-one-condition-out test also showed that the calibration transferred reasonably from clear, night, rain and construction data to the held-out condition, with an expected calibration

error of between 0.041 and 0.067. At 0.083, it was an exception to the occlusion data, and thus direct occlusion coverage was required during calibration. The above sensitivity analyses support the reported trends and indicate areas for re-estimation.

Conclusion

This paper establishes an index for the transparency of deep autonomous driving networks by combining semantic alignment, causal faithfulness, temporal consistency, calibrated confidence and operational cost. Connect the multi-task driving representations to road-scene concepts in the proposed framework, validate these concepts with plausible counterfactuals, and suppress explanation confidence when the predictive evidence is uncertain. Across route-disjoint tests and blinded reviews, the framework maintained the perception-and-steering quality of the original and provided explanations that were more faithful, stable and useful than those generated by conventional post-hoc methods.

The extent and detail of the concept bank are still limited, the realism of the generated counterfactual scenes is not high, and it is difficult to guarantee temporal accuracy due to unreliable motion estimation. Some crowded or reflective scenes still have semantically incomplete descriptions, and the experiments do not consider every sensor suite, geography, or behaviour policy in urban driving. Transparency scores should thus support, rather than replace, scenario testing, redundancy, formal safety analysis and expert event reconstruction.

Future work will learn expandable concept vocabularies from multimodal supervision, generate counterfactuals under explicit three-dimensional and traffic-rule constraints, and link explanation confidence to fleet-scale drift monitoring. Add evaluation of sensor fusion and interactive planning to assess how explanations affect engineering decisions over time through long-term field studies. A general benchmark of task performance, explanation validity, reviewer behaviour and runtime could be introduced to enhance the comparability and general applicability of transparency claims in safety assurance.

Author Contributions

Lidia Kaczmarowa contributes to conceptualization, methodology, software, validation, analysis, investigation, data collection, draft preparation, manuscript editing, visualization, supervision, project administration, and funding acquisition. Oliwia Barbara Budzyńska contributes to validation, analysis, investigation, data collection, draft preparation. All authors have read and agreed with the manuscript before its submission and publication.

Funding

This research received no specific financial support from any funding agency.

Institutional Review Board Statement

Not applicable.

References

[1] van der Velden, B. H. M., Kuijf, H. J., Gilhuijs, K. G. A., & Viergever, M. A. (2022). Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. Medical image analysis, 79, 102470. https://doi.org/10.1016/j.media.2022.102470

[2] Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. The Lancet. Digital health, 3(11), e745-e750. https://doi.org/10.1016/S2589-7500(21)00208-9

[3] Cho, Y. R., & Kang, M. (2020). Interpretable machine learning in bioinformatics. Methods (San Diego, Calif.), 179, 1-2. https://doi.org/10.1016/j.ymeth.2020.05.024

[4] Rudin, C. (2019). Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. Nature machine intelligence, 1(5), 206-215. https://doi.org/10.1038/s42256-019-0048-x

[5] Watson, D. S. (2022). Interpretable machine learning for genomics. Human genetics, 141(9), 1499-1513. https://doi.org/10.1007/s00439-021-02387-9

[6] Rasheed, K., Qayyum, A., Ghaly, M., Al-Fuqaha, A., Razi, A., & Qadir, J. (2022). Explainable, trustworthy, and ethical machine learning for healthcare: A survey. Computers in biology and medicine, 149, 106043. https://doi.org/10.1016/j.compbiomed.2022.106043

[7] Qiu, W., Chen, H., Dincer, A. B., Lundberg, S., Kaeberlein, M., & Lee, S. I. (2022). Interpretable machine learning prediction of all-cause mortality. Communications medicine, 2, 125. https://doi.org/10.1038/s43856-022-00180-x

[8] Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. Wiley interdisciplinary reviews. Data mining and knowledge discovery, 9(4), e1312. https://doi.org/10.1002/widm.1312

[9] Garbulowski, M., Diamanti, K., Smolińska, K., Baltzer, N., Stoll, P., Bornelöv, S., …, Komorowski, J. (2021). R.ROSETTA: an interpretable machine learning framework. BMC bioinformatics, 22(1), 110. https://doi.org/10.1186/s12859-021-04049-z

[10] Hauser, K., Kurz, A., Haggenmüller, S., Maron, R. C., von Kalle, C., Utikal, J. S., …, Brinker, T. J. (2022). Explainable artificial intelligence in skin cancer recognition: A systematic review. European journal of cancer (Oxford, England : 1990), 167, 54-69. https://doi.org/10.1016/j.ejca.2022.02.025

[11] Yang, C. C. (2022). Explainable Artificial Intelligence for Predictive Modeling in Healthcare. Journal of healthcare informatics research, 6(2), 228-239. https://doi.org/10.1007/s41666-022-00114-1

[12] Wilhelm, D., Bouarfa, L., Navab, N., Meining, A., Müller-Stich, B., Jarc, A., & Padoy, N. (2020). Artificial Intelligence in Visceral Medicine. Visceral medicine, 36(6), 471-475. https://doi.org/10.1159/000512440

[13] Sidak, D., Schwarzerová, J., Weckwerth, W., & Waldherr, S. (2022). Interpretable machine learning methods for predictions in systems biology from omics data. Frontiers in molecular biosciences, 9, 926623. https://doi.org/10.3389/fmolb.2022.926623

[14] Azodi, C. B., Tang, J., & Shiu, S. H. (2020). Opening the Black Box: Interpretable Machine Learning for Geneticists. Trends in genetics : TIG, 36(6), 442-455. https://doi.org/10.1016/j.tig.2020.03.005

[15] Schinle, M., Erler, C., Hess, M., & Stork, W. (2022). Explainable Artificial Intelligence in Ambulatory Digital Dementia Screenings. Studies in health technology and informatics, 294, 123-124. https://doi.org/10.3233/SHTI220411

[16] Thomas, A. W., Ré, C., & Poldrack, R. A. (2022). Interpreting mental state decoding with deep learning models. Trends in cognitive sciences, 26(11), 972-986. https://doi.org/10.1016/j.tics.2022.07.003

[17] Lumbreras, S. (2022). The Synergies Between Understanding Belief Formation and Artificial Intelligence. Frontiers in psychology, 13, 868903. https://doi.org/10.3389/fpsyg.2022.868903

[18] Ladbury, C., Zarinshenas, R., Semwal, H., Tam, A., Vaidehi, N., Rodin, A. S., …, Amini, A. (2022). Utilization of model-agnostic explainable artificial intelligence frameworks in oncology: a narrative review. Translational cancer research, 11(10), 3853-3868. https://doi.org/10.21037/tcr-22-1626

[19] Fellous, J. M., Sapiro, G., Rossi, A., Mayberg, H., & Ferrante, M. (2019). Explainable Artificial Intelligence for Neuroscience: Behavioral Neurostimulation. Frontiers in neuroscience, 13, 1346. https://doi.org/10.3389/fnins.2019.01346

[20] Farrell, S., Mitnitski, A., Rockwood, K., & Rutenberg, A. D. (2022). Interpretable machine learning for high-dimensional trajectories of aging health. PLoS computational biology, 18(1), e1009746. https://doi.org/10.1371/journal.pcbi.1009746

[21] Payrovnaziri, S. N., Chen, Z., Rengifo-Moreno, P., Miller, T., Bian, J., Chen, J. H., …, He, Z. (2020). Explainable artificial intelligence models using real-world electronic health record data: a systematic scoping review. Journal of the American Medical Informatics Association : JAMIA, 27(7), 1173-1185. https://doi.org/10.1093/jamia/ocaa053

[22] Hrovatin, K., Fischer, D. S., & Theis, F. J. (2022). Toward modeling metabolic state from single-cell transcriptomics. Molecular metabolism, 57, 101396. https://doi.org/10.1016/j.molmet.2021.101396

[23] Wells, L., & Bednarz, T. (2021). Explainable AI and Reinforcement Learning-A Systematic Review of Current Approaches and Trends. Frontiers in artificial intelligence, 4, 550030. https://doi.org/10.3389/frai.2021.550030

[24] Loh, H. W., Ooi, C. P., Seoni, S., Barua, P. D., Molinari, F., & Acharya, U. R. (2022). Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011-2022). Computer methods and programs in biomedicine, 226, 107161. https://doi.org/10.1016/j.cmpb.2022.107161

[25] Tang, R., Liu, N., Yang, F., Zou, N., & Hu, X. (2022). Defense Against Explanation Manipulation. Frontiers in big data, 5, 704203. https://doi.org/10.3389/fdata.2022.704203

[26] Gimeno, M., San José-Enériz, E., Villar, S., Agirre, X., Prosper, F., Rubio, A., & Carazo, F. (2022). Explainable artificial intelligence for precision medicine in acute myeloid leukemia. Frontiers in immunology, 13, 977358. https://doi.org/10.3389/fimmu.2022.977358

[27] Ning, Y., Ong, M. E. H., Chakraborty, B., Goldstein, B. A., Ting, D. S. W., Vaughan, R., & Liu, N. (2022). Shapley variable importance cloud for interpretable machine learning. Patterns (New York, N.Y.), 3(4), 100452. https://doi.org/10.1016/j.patter.2022.100452

[28] Thomas, J., Ledger, G. A., & Mamillapalli, C. K. (2020). Use of artificial intelligence and machine learning for estimating malignancy risk of thyroid nodules. Current opinion in endocrinology, diabetes, and obesity, 27(5), 345-350. https://doi.org/10.1097/MED.0000000000000557

[29] Zhang, X., Tian, Y., Chen, L., Hu, X., & Zhou, Z. (2022). Machine Learning: A New Paradigm in Computational Electrocatalysis. The journal of physical chemistry letters, 13(34), 7920-7930. https://doi.org/10.1021/acs.jpclett.2c01710

[30] Ascher, S., Wang, X., Watson, I., Sloan, W., & You, S. (2022). Interpretable machine learning to model biomass and waste gasification. Bioresource technology, 364, 128062. https://doi.org/10.1016/j.biortech.2022.128062

[31] Tan, R., Gao, L., Khan, N., & Guan, L. (2022). Interpretable Artificial Intelligence through Locality Guided Neural Networks. Neural networks : the official journal of the International Neural Network Society, 155, 58-73. https://doi.org/10.1016/j.neunet.2022.08.009

[32] Fassio, A. V., Shub, L., Ponzoni, L., McKinley, J., O'Meara, M. J., Ferreira, R. S., …, de Melo Minardi, R. C. (2022). Prioritizing Virtual Screening with Interpretable Interaction Fingerprints. Journal of chemical information and modeling, 62(18), 4300-4318. https://doi.org/10.1021/acs.jcim.2c00695

[33] Sun, Y., Song, Q., & Liang, F. (2022). Consistent Sparse Deep Learning: Theory and Computation. Journal of the American Statistical Association, 117(540), 1981-1995. https://doi.org/10.1080/01621459.2021.1895175

[34] Alderden, J., Kennerly, S. M., Wilson, A., Dimas, J., McFarland, C., Yap, D. Y., …, Yap, T. L. (2022). Explainable Artificial Intelligence for Predicting Hospital-Acquired Pressure Injuries in COVID-19-Positive Critical Care Patients. Computers, informatics, nursing : CIN, 40(10), 659-665. https://doi.org/10.1097/CIN.0000000000000943

[35] Xie, F., Ning, Y., Yuan, H., Goldstein, B. A., Ong, M. E. H., Liu, N., & Chakraborty, B. (2022). AutoScore-Survival: Developing interpretable machine learning-based time-to-event scores with right-censored survival data. Journal of biomedical informatics, 125, 103959. https://doi.org/10.1016/j.jbi.2021.103959

[36] Andrienko, N., Andrienko, G., Adilova, L., Wrobel, S., & Rhyne, T. M. (2022). Visual Analytics for Human-Centered Machine Learning. IEEE computer graphics and applications, 42(1), 123-133. https://doi.org/10.1109/MCG.2021.3130314

[37] Kavvas, E. S., Yang, L., Monk, J. M., Heckmann, D., & Palsson, B. O. (2020). A biochemically-interpretable machine learning classifier for microbial GWAS. Nature communications, 11(1), 2580. https://doi.org/10.1038/s41467-020-16310-9

[38] Simon, S. T., Trinkley, K. E., Malone, D. C., & Rosenberg, M. A. (2022). Interpretable Machine Learning Prediction of Drug-Induced QT Prolongation: Electronic Health Record Analysis. Journal of medical Internet research, 24(12), e42163. https://doi.org/10.2196/42163