Mesopotamian journal of Big Data
Vol. **(2025)**, 2025, **pp**. 448–475
DOI: https://doi.org/10.58496/MJBD/2025/027 ISSN: 2958-6453
https://mesopotamian.press/journals/index.php/BigData

Research Article

Validation Framework for Robust and Explainable Machine Learning Models in Autism Spectrum Disorder Triage

Ghadeer Ghazi Shayea [1,2], Mohd Hazli Mohammed Zabil [1], A. S. Albahri [2,3,*], Rula A. Hamid [4], O. S. Albahri [5], A. H. Alamoodi [6,7], Iman Mohamad Sharaf [8]

1 Universiti Tenaga Nasional (UNITEN), Kajang, Malaysia
2 Technical Engineering College, Imam Ja'afar Al-Sadiq University, Baghdad, Iraq
3 University of Information Technology and Communications (UoITC), Baghdad, Iraq
4 College of Business Informatics, University of Information Technology and Communications (UOITC), Baghdad, Iraq
5 Institute of Innovation, Science and Sustainability (IISS), Federation University, Melbourne Campus, Australia
6 Applied Science Research Center, Applied Science Private University, Amman, Jordan
7 GUST Engineering and Applied Innovation Research Center (GEAR), Gulf University for Science and Technology, Mishref, Kuwait
8 Department of Basic Sciences, Higher Technological Institute, Tenth of Ramadan City, Egypt

ARTICLEINFO

Article History
Received: 07 Sep 2025
Revised: 24 Oct 2025
Accepted: 16 Nov 2025
Published: 21 Nov 2025

Keywords
Autism Spectrum Disorder (ASD)
Triage
Robust Machine Learning
Adversarial Attacks
Explainable Artificial Intelligence
(XAI)
Fuzzy Multi-Criteria Decision- Making (MCDM)

ABSTRACT

The development of real-time triage applications for Autism Spectrum Disorder (ASD) is a critical challenge due to the rising prevalence of ASD and the urgent need for efficient resource allocation in healthcare systems. Previous studies have applied machine learning (ML) to ASD triage; however, most approaches overlook robustness against adversarial attacks, provide limited benchmarking across multiple evaluation criteria, and lack explainability to support clinical adoption. Building on our earlier work, which introduced a fuzzy evaluation and benchmarking framework using the 2-Tuple Linguistic Fermatean Fuzzy Decision by Opinion Score Method (2TLFFDOSM), this study proposes a comprehensive five-stage validation and evaluation framework. The framework systematically validates fuzzy-based rankings against raw performance metrics, conducts dual-perspective analysis under normal and adversarial conditions, performs sensitivity analysis across ten weighting scenarios, and integrates explainable AI (LIME, PFI, Integrated Gradients, and PDP) to interpret feature contributions before and after adversarial perturbations. Finally, a checklist benchmarking approach is used to position the framework against five recent studies.

1. INTRODUCTION

Autism Spectrum Disorder (ASD) has become one of the most pressing global health concerns, with rising prevalence rates creating increasing demand for timely and accurate triage in clinical environments [1]. Triage systems for ASD aim to prioritize patients into urgent, moderate, or minor categories to optimize healthcare resources and ensure rapid intervention for critical cases [2]. Traditional triage processes are often time-consuming and heavily dependent on expert judgment, which can introduce bias and reduce scalability. As a result, the integration of machine learning (ML) techniques has emerged as a promising solution to enhance the accuracy, efficiency, and reliability of ASD triage [3].

*Corresponding author. Email: ahmed.bahri1978@gmail.com

While numerous studies have investigated the application of ML for ASD diagnosis and triage, most approaches remain limited in three critical respects [4][5]. First, they primarily emphasize predictive accuracy without adequately addressing robustness against adversarial attacks, despite the vulnerability of ML models to malicious perturbations in sensitive healthcare contexts. Second, the majority of studies lack a comprehensive benchmarking framework to fairly evaluate and compare competing ML models across multiple performance criteria. Third, explainability is often overlooked, leaving end-users—particularly clinicians—without insight into why a given model produces specific triage outcomes. These limitations hinder both the trustworthiness and adoption of ML-driven triage systems in real-world clinical practice.

To address these issues, our previous study [3] introduced a fuzzy evaluation and benchmarking framework that fused medical and sociodemographic features from 1,296 ASD patients, applied Principal Component Analysis (PCA) for dimensionality reduction, and evaluated eight ML algorithms under both normal and adversarial test conditions. A novel 2-Tuple Linguistic Fermatean Fuzzy Decision by Opinion Score Method (2TLFFDOSM) was proposed to benchmark models against 18 criteria, integrating expert judgment through external aggregation. The findings revealed that Logistic Regression (LR) demonstrated the highest robustness under adversarial settings, while Support Vector Machine (SVM) excelled under normal conditions. Despite these contributions, the study left open several critical gaps: (i) the absence of systematic validation of the fuzzy-based rankings against raw performance metrics, (ii) no sensitivity analysis to assess the stability of results under varying criteria weights, and (iii) the lack of model explainability to justify and interpret predictions.

Building upon this foundation, the present study develops and validates a five-stage evaluation and validation framework to overcome the aforementioned limitations. Specifically, the contributions of this paper are as follows:

1. Systematic Ranking Validation – Cross-verification of fuzzy-based rankings with normalized raw performance metrics to ensure consistency and reliability.
2. Dual-Perspective Evaluation – Independent and comparative assessment of ML model rankings under both normal and adversarial conditions.
3. Sensitivity Analysis – Examination of ranking stability under ten distinct weighting scenarios to account for expert bias and uncertainty.
4. Explainability Analysis – Deployment of multiple XAI tools (LIME, PFI, Integrated Gradients, and PDP) to interpret predictions and highlight feature importance both before and after adversarial perturbations.
5. Checklist Benchmarking – Comparative analysis against five recent studies using a structured set of theoretical- and application-level checklists to demonstrate the novelty and comprehensiveness of the proposed framework.

By systematically validating, stress-testing, and explaining ML model selection for ASD triage, this study advances the state of the art in healthcare-focused ML benchmarking. The proposed framework not only strengthens the methodological foundations of our earlier work but also provides clinicians and researchers with a transparent, reproducible, and security-aware decision-making pipeline suitable for deployment in real-world medical environments.

2. FRAMEWORK VALIDATION AND EVALUATION

This section presents the validation and evaluation framework for the proposed methodology, as presented in previous work [3]. In Section 2.1, we summarise the previously published work focused on the development of a fuzzy evaluation and benchmarking framework for robust machine learning models in real-time autism triage applications. The presented study then developed four assessment processes. Initially, the systematic ranking of the ML models is evaluated in Section 2.2, with a focus on their ranking results. Subsequently, the analysis examines the impact of changing the criteria weight on the ranking result, utilizing the sensitivity analysis presented in Section 2.3. Section 2.4 involves determining the interpretability of the selected ML models, shedding light on their decision-making processes. Finally, the proposed framework is compared using a checklist benchmarking approach, underscoring its strengths and potential areas for improvement, and is presented in Section 2.5.

2.1. Previous Methodology: Results and Limitations

The previous study published in [3] addresses the challenge of building robust machine learning (ML) models for autism spectrum disorder (ASD) triage, where patients are categorised into urgent, moderate, or minor groups. The central

issues tackled include model robustness, selection, and the ability to perform effectively under both normal and adversarial attack scenarios.

The study proposed a three-phase framework. In the first phase, dataset identification and preprocessing were conducted using records from 1296 ASD patients with 19 medical and sociodemographic features. Labels were categorised as urgent, moderate, and minor, and dimensionality reduction was achieved through Principal Component Analysis (PCA), which reduced the features to 12 principal components while retaining 79.6% of the dataset's variance. The second phase involved the development and testing of eight supervised ML algorithms—logistic regression (LR), decision tree (DT), neural network (NN), stochastic gradient descent (SGD), random forest (RF), k-nearest neighbors (kNN), naïve Bayes (NB), and support vector machine (SVM). These models were evaluated across nine performance metrics, including accuracy, F1-score, precision, recall, log loss, and specificity, under both normal and adversarial conditions. Adversarial samples were generated using the Fast Gradient Sign Method (FGSM) to simulate attacks. In the third phase, a Decision Matrix (DM) comprising 18 criteria (nine from normal conditions and nine from adversarial tests) was established. To benchmark the models, the authors introduced the novel 2-Tuple Linguistic Fermatean Fuzzy Decision by Opinion Score Method (2TLFFDOSM), which effectively managed conflicting criteria, importance weighting, and uncertainty. External group aggregation was also employed to integrate expert opinions.

The findings revealed that LR consistently achieved the highest robustness, ranking first in the 2TLFFDOSM evaluation with a score of 1.3370, followed by SVM and RF. The developed methodology in [3] makes three major contributions: (1) the creation of the first ASD triage dataset integrating medical and sociodemographic features with PCA-based fusion; (2) the development of the novel 2TLFFDOSM fuzzy method for multi-criteria evaluation and benchmarking of ML models; and (3) the establishment of a comprehensive framework for robust real-time ASD triage.

Despite these contributions, the study acknowledges several limitations. The methodology lacked explainability features, highlighting the need for future integration of interpretable AI methods. This motivated the authors to develop the current framework for validation and evaluation of the previous results, as presented in the next sections.

2.2. Systematic Ranking

In this section, a systematic ranking assessment is conducted to evaluate the outputs of the ML models. This procedure draws from earlier MCDM research [6][7]. To validate the results obtained from the 2TLF-FDOSM-based ranking [3], a direct comparison was made with the raw performance metrics of the developed ML models. This approach ensures that the ranking produced by the MCDM method aligns with the actual predictive performance of the models. Accordingly, the benchmarking validation was structured around the following procedure:

1. **Performance Metric Evaluation**: Each ML model was evaluated based on its core performance metrics with the consideration of Perspective 1 Criteria (Metrics for ML Models based on Normal test examples) and Perspective 2 Criteria (Metrics for ML Models based on Adversarial attack examples).
2. **Min-Max Normalization:** All metric values were normalized using the min-max scaling method, as defined by the following equation:

$$Normalized\ Score = \frac{\text{Max}-\text{OriginalValue}}{\text{Max}-Min} \quad (1)$$

where Original Value refers to the raw score of the metric.

This approach scales metric values into a uniform range, where lower normalized values represent better performance in cases such as training and testing time, and higher values reflect better predictive quality for metrics like accuracy or F1 Score.

3. **Grouping and Statistical Aggregation**: All ML models were divided into four equally sized groups based on their normalized scores to maintain a balanced comparative framework. For each group, the mean and standard deviation (STD) were calculated to measure the average performance and internal variability within clusters.

4. **Ranking Comparison:** A reference ranking was derived from the normalized metric scores and compared with the 2TLF-FDOSM crisp score ranking [3]. This comparison aimed to assess the consistency and reliability of the fuzzy decision-making method in capturing true model performance.
5. **Validation Logic**: The comparison follows the 2TLF-FDOSM validation logic, where the group with the lowest mean value is considered the most valid and accurate representation of top performance. The validation is confirmed if the following conditions are met:
    - The mean of the first group is lower than that of the second group.
    - The mean of the second group is higher than that of the first, ensuring a valid progressive performance scale.

If the models ranked highest using normalized performance metrics (e.g., LR, SVM) match those ranked top by 2TLF-FDOSM, it supports the validity and robustness of the fuzzy-based MCDM approach in evaluating ML model quality. This structured validation procedure confirms the credibility of the ranking outcomes derived through 2TLF-FDOSM.

2.3. Sensitivity Analysis

By generating a series of filter selections and utilizing them to make decisions among the composite alternatives, MCDM reduces bias. By conducting a sensitivity analysis, the impact of criterion weights on the decision-making process is ascertained. Its purpose is to assess the consistency of the ranking with respect to the criterion weights. By generating a series of filter selections and utilizing them to make decisions among the composite alternatives, MCDM reduces bias. By conducting a sensitivity analysis, the impact of criterion weights on the decision-making process is ascertained. Its purpose is to assess the consistency of the ranking with respect to the criterion weights [8]. As a result, several MCDM techniques use various mathematical procedures, producing unique answers to the same issue. Sensitivity analysis therefore entails decision-making due to unclear input values [9][10] in order to demonstrate the validity of the evaluation and benchmarking choice reported in earlier work [3]. The new weights of the criteria within the two viewpoints are changed to either increase or decrease through a series of phases in the suggested evaluation process. The two perspective criteria are given % values in order to add the criteria weights. For the weight percentage M, seven options were suggested in this study: 50%:50%, 60%:40%, 70%:30%, 80%:20%, 40%:60%, 30%:70%, and 20%:80%. The sensitive weights are calculated as follows:

$$Ws = \left(\frac{M}{N}\right) / 100 \tag{2}$$

where:

Ws: sensitivity weights for each perspective,

M: Weight percentage,

N: Number of criteria for each perspective

2.4. Explainability Analysis

Processes have been incorporated into the XAI model to facilitate understanding and confidence in the outcomes and outputs produced by the benchmarked ML model that was selected in the study of [3]. During this stage, the XAI model was employed to describe the most precise ML model for the ASD fused dataset, as determined by the 2TLF-FDOSM outcome among the eight ML algorithms, utilizing the LIME technique. LIME endeavours to comprehend the ML model in an unbiased manner by manipulating the input of data samples and scrutinizing the resulting fluctuations in predictions. The LIME model supplies a local approximation of the inference point to facilitate comprehension of the inferences generated by the robust ML model. The algorithm constructs a linear regression in the vicinity of a particular inference point that requires elucidation. When coping with difficult-to-interpret models, such as black-box models or models with a large number of features, it is advantageous to employ this technique.

LIME enhances trust in predictions and increases the transparency of machine learning models by offering understandable explanations for individual predictions [11]. The following points summarize the functioning of works [12]:

1. The specific instance (i.e., the data point) requiring explanation is chosen.
2. A collection of altered instances is created by randomly adjusting the features of the original instance.

3. The complex model is utilized to produce predictions for all the altered instances.
4. A surrogate model (for example, a linear regression model) is trained to estimate the complex model's output based on the altered instances and their respective predictions.
5. The key features are determined by examining the coefficients of the surrogate model.
6. An explanation is formed by emphasizing the most significant features and their contributions to the prediction.

2.5. Checklist Comparison Approach

In this section, a thorough comparison of the previously established methodology in [3] is performed against the literature using a checklist benchmarking approach. A commonly utilized comparison methodology found in the literature is checklist benchmarking. This method entails assessing various significant checklists presented as factors to highlight the uniqueness of the current work. The aspects considered include: Normal/Adversarial Perspectives, Fusion Improvements, the Development of the MCDM Selection Method, Decision Matrix Development, Medical and Sociodemographic Features, issues with ML criteria, and the ML explainability process. This examination highlights the distinctive contributions and potential areas for enhancement in the suggested methodology.

3. RESULTS AND DISCUSSION

This section provides a comprehensive validation and evaluation of the proposed 2TLF-FDOSM-based decision-making framework for ranking ML models under varying input conditions. The evaluation begins with a systematic ranking of the results obtained from a previous study [3] and presented in Section 3.1. Then, another ranking analysis for the model's performance under clean (normal) data is presented in Section 3.2, followed by an analysis of adversarial input in Section 3.3. A hybrid input ranking analysis combining normal and adversarial perspectives is presented in Section 3.4. Each of these sections also includes a dedicated sensitivity analysis to assess the robustness of the rankings against changes in criteria weights. Section 3.5 focuses on the explainability of the selected top-ranked model, LR, using advanced interpretability tools.

3.1. Systematic Ranking

Table 1 presents the validation results based on the normalized performance metrics for all ML models. The models were grouped into four equally sized clusters (G1 to G4), and for each group, the mean ± STD of the normalized values was calculated to assess average performance and internal consistency.

TABLE I. VALIDATION OF GROUP RESULTS.

Alternative ID	ML Models	Order	ED	Mean ± STD
A8	LR	1	G1	0.3577 ± 0.3492
A3	SVM	2		
A5	RF	3	G2	0.4080 ± 0.3306
A6	NN	4		
A7	NB	5	G3	0.5378 ± 0.3545
A2	DT	6		
A1	kNN	7	G4	0.6516 ± 0.3726
A4	SGD	8		

The validation logic follows the principle that lower group mean values indicate stronger model performance. The results show the following trend:

- **G1 (LR, SVM)** has the **lowest mean value** (0.3577), indicating it contains the **top-performing models**.
- **G2 (RF, NN)** shows slightly higher performance metrics, still within an acceptable high-performance range.
- **G3 (NB, DT)** and **G4 (kNN, SGD)** show progressively higher means, confirming **lower predictive performance** relative to G1.

The increasing order of group means (G1 < G2 < G3 < G4) supports the validity of the ranking produced by the 2TLF-FDOSM method. This progressive structure validates that the ranking is not random, but systematically aligned with the actual performance characteristics of the ML models. The normalized raw performance metrics strongly validate the ranking generated by 2TLF-FDOSM. The groupings show a clear distinction in performance levels, and the ranking order

is consistent with statistical measures, thereby reinforcing the robustness and reliability of the proposed MCDM-based evaluation framework.

3.2. Normal-Only Ranking

This section presents the ranking results of eight ML models based solely on clean (normal) input data, using the 2TLF-FDOSM methodology. Rankings were derived from expert opinion matrices reflecting Perspective 1, which evaluates model performance exclusively under non-adversarial conditions.

The experts assessed the models using criteria outlined in [3], including accuracy, precision, recall, F1-score, training time, and testing time. Each expert's scores were independently processed through the 2TLF-FDOSM to yield individual rankings, followed by external aggregation to determine final ranks. The results are shown in Table 2.

TABLE II. 2TLF-FDOSM SCORES AND RANKINGS FOR ML MODELS (NORMAL DATA ONLY)

		2TLF-FDOSM Scores and corresponding ranks							
ML Models		**Expert 1**	**Rank**	**Expert 2**	**Rank**	**Expert 3**	**Rank**	**External Aggregation**	**Rank**
A1	**kNN**	1.0890	6	1.0160	8	0.9669	7	1.0266	7
A2	**DT**	1.0472	7	1.1652	6	1.0180	6	1.0825	6
A3	**SVM**	1.5205	1	1.5078	1	1.0570	5	1.4057	1
A4	**SGD**	0.9970	8	1.0382	7	0.9329	8	0.9914	8
A5	**RF**	1.3488	2	1.4591	2	1.1833	3	1.3443	2
A6	**NN**	1.2385	4	1.3625	3	1.4106	1	1.3427	3
A7	**NB**	1.1088	5	1.1853	5	1.1517	4	1.1508	5
A8	**LR**	1.2834	3	1.3441	4	1.2638	2	1.2983	4

SVM (A3) ranked first under normal data, reflecting its strong generalization ability and consistently high classification performance. RF (A5) and NN (A6) followed closely, with all three models demonstrating robust scores across experts. Notably, NN (A6) was ranked highest by Expert 3, but placed third overall due to variability in other experts' assessments. LR (A8) showed a solid all-around performance, ranking fourth, likely due to its computational efficiency and reliable baseline metrics. NB (A7) is placed in the mid-tier range, balancing speed with moderate predictive strength.

The lowest performers—SGD (A4), kNN (A1), and DT (A2)—struggled to achieve competitive rankings. These results may be attributed to instability on clean data (SGD), sensitivity to noisy feature distributions (kNN), and overfitting tendencies (DT). This ranking demonstrates how structured MCDM evaluations like 2TLF-FDOSM can integrate subjective expertise into quantitative performance assessments under standard operating conditions.

3.3. Adversarial-Only Ranking

This section presents the ranking of the same ML models under adversarial conditions, using the 2TLF-FDOSM based on Perspective 2, which focuses on model behavior when exposed to perturbed or maliciously crafted input data. The criteria for evaluation remain identical, but the performance metrics are recalculated based on adversarial test inputs. Expert scores were processed similarly to the previous section, producing both individual and aggregated rankings shown in Table 3.

TABLE III. 2TLF-FDOSM SCORES AND RANKINGS FOR ML MODELS (ADVERSARIAL DATA ONLY)

		2TLF-FDOSM Scores and corresponding ranks							
ML Models		**Expert 1**	**Rank**	**Expert 2**	**Rank**	**Expert 3**	**Rank**	**External Aggregation**	**Rank**
A1	**kNN**	1.0554	5	1.0468	4	0.9761	8	1.0268	6
A2	**DT**	0.9396	8	0.9860	7	1.1030	5	1.0186	7
A3	**SVM**	1.0246	6	1.0180	6	1.4293	1	1.2026	3
A4	**SGD**	1.0132	7	0.8975	8	0.9901	7	0.9695	8
A5	**RF**	1.1853	4	1.0329	5	1.4092	2	1.2348	2

A6	**NN**	1.1871	3	1.1080	3	1.1665	4	1.1548	5
A7	**NB**	1.2834	2	1.2568	2	1.0160	6	1.1998	4
A8	**LR**	1.4909	1	1.3927	1	1.1877	3	1.3726	1

A8 emerged as the top-performing model under adversarial attack, maintaining consistently high ranks from all experts. Despite its simplicity, LR demonstrated stable behavior under perturbation—possibly due to its linear decision boundary and lower susceptibility to adversarial gradient shifts.

RF (A5) placed second, benefiting from ensemble robustness across diverse decision trees. SVM (A3), previously ranked first under normal conditions, dropped to third—indicating some vulnerability to adversarial manipulation, despite retaining high confidence from one expert.

NB (A7) and Neural Network (A6) occupied mid-level ranks. NB's relative simplicity may have limited the surface for adversarial exploitation, while NN showed mixed robustness due to its complexity and sensitivity to crafted input distortions. The bottom performers—SGD (A4), DT (A2), and kNN (A1)—showed significant degradation in adversarial settings. SGD's rank remained consistently low, likely due to its reliance on gradient stability, while DT and kNN are known to be brittle when small input changes lead to major output shifts.

Comparing the normal-only and adversarial-only rankings clearly reveals significant variation in model performance across the two scenarios. Models such as SVM and LR changed positions dramatically, underscoring the contextual dependency of model effectiveness. These results validate a central argument of this study: *The evaluation and benchmarking of ML models must account for both clean and adversarial scenarios simultaneously. A single, one-dimensional evaluation is insufficient, as performance under standard conditions does not reliably predict robustness under attack*. The dual-perspective rankings provided by 2TLF-FDOSM highlight performance variation and encourage a more comprehensive and security-aware approach to model selection in real-world applications.

3.4. Sensitivity Analysis

The application of MCDM methods inherently involves weighting various evaluation criteria, which introduces subjectivity and potential bias. Sensitivity analysis helps mitigate this by examining the effect of varying these weights on the final rankings, thereby assessing the robustness and reliability of the decision-making process. In this study, sensitivity analysis was employed to validate the stability of the ML model rankings derived from the 2TLF-FDOSM approach under two evaluation perspectives: Perspective 1, involving clean (normal) test examples, and Perspective 2, involving adversarial attack examples.

This analysis examines how varying weighting schemes across evaluation criteria impact the model rankings. Following the guidance of previous works [8][9][10], ten distinct weighting scenarios were proposed, as detailed in Table 4. The first scenario (S1) assigns equal weights to all criteria, while the remaining scenarios vary the weight of one criterion each to simulate potential expert emphasis shifts.

TABLE IV. SENSITIVITY SCENARIOS WITH CRITERIA WEIGHTS

Criteria		Sensitivity Scenarios Weights									
		S1: Equal weights	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**S9**	**S10**
C1	**train time**	0.1111	0.2	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
C2	**test time**	0.1111	0.1	0.2	0.1	0.1	0.1	0.1	0.1	0.1	0.1
C3	**AUC**	0.1111	0.1	0.1	0.2	0.1	0.1	0.1	0.1	0.1	0.1
C4	**CA**	0.1111	0.1	0.1	0.1	0.2	0.1	0.1	0.1	0.1	0.1
C5	**F1**	0.1111	0.1	0.1	0.1	0.1	0.2	0.1	0.1	0.1	0.1
C6	**Precision**	0.1111	0.1	0.1	0.1	0.1	0.1	0.2	0.1	0.1	0.1
C7	**Recall**	0.1111	0.1	0.1	0.1	0.1	0.1	0.1	0.2	0.1	0.1
C8	**LogLoss**	0.1111	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.2	0.1
C9	**Specificity**	0.1111	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.2

The modified weight scenarios were applied independently to both perspectives, and the resulting ranks were compared with the original 2TLF-FDOSM rankings to observe fluctuations.

- **Sensitivity Results: Perspective 1 (Normal Data):** Table 5 and Figure 1 illustrate the model ranking variations under different weighting scenarios for clean test data.

TABLE V. SENSITIVITY ANALYSIS RANKS FOR ML MODELS (NORMAL TEST EXAMPLES)

ML Models		Ranks										
		Original Rank	S1: Equal Weights	S2	S3	S4	S5	S6	S7	S8	S9	S10
A1	**kNN**	7	7	7	8	7	7	7	7	7	7	7
A2	**DT**	6	6	6	6	6	6	6	6	6	6	6
A3	**SVM**	1	1	1	1	1	1	1	1	1	1	1
A4	**SGD**	8	8	8	7	8	8	8	8	8	8	8
A5	**RF**	2	2	2	3	3	2	2	3	2	3	2
A6	**NN**	3	3	3	2	2	3	3	2	3	2	3
A7	**NB**	5	5	5	5	5	5	5	5	5	5	5
A8	**LR**	4	4	4	4	4	4	4	4	4	4	4

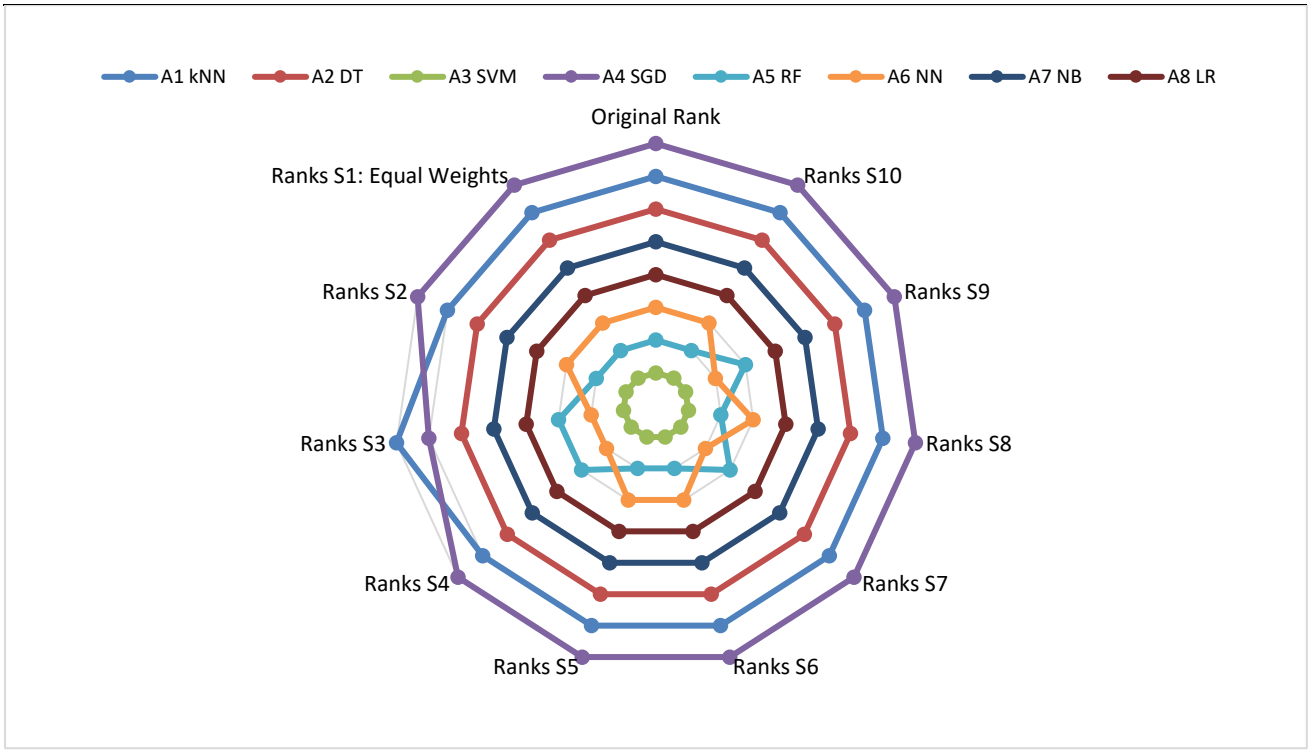

Figure 1. Rank variation under different weights for Perspective 1 (Normal Data)

The rankings for SVM, RF, and NN remained largely stable, indicating strong robustness of these models under clean data conditions across a wide range of criteria emphasis. SGD and kNN showed minor variability, reinforcing their previously observed limitations under normal evaluation.

- **Sensitivity Results: Perspective 2 (Adversarial Data):** Table 6 and Figure 2 show how rankings shifted for the same models under adversarial test scenarios.

TABLE VI. SENSITIVITY ANALYSIS RANKS FOR ML MODELS (ADVERSARIAL EXAMPLES)

ML Models		Ranks										
		Original Rank	S1: Equal weights	S2	S3	S4	S5	S6	S7	S8	S9	S10
A1	kNN	6	6	6	8	6	6	7	6	7	6	6
A2	DT	7	7	8	6	7	7	6	7	6	7	7
A3	SVM	3	3	4	4	4	3	3	3	3	2	3
A4	SGD	8	8	7	7	8	8	8	8	8	8	8
A5	RF	2	2	2	2	3	2	2	2	2	3	2
A6	NN	5	5	5	5	5	5	5	5	5	5	5
A7	NB	4	4	3	3	2	4	4	4	4	4	4
A8	LR	1	1	1	1	1	1	1	1	1	1	1

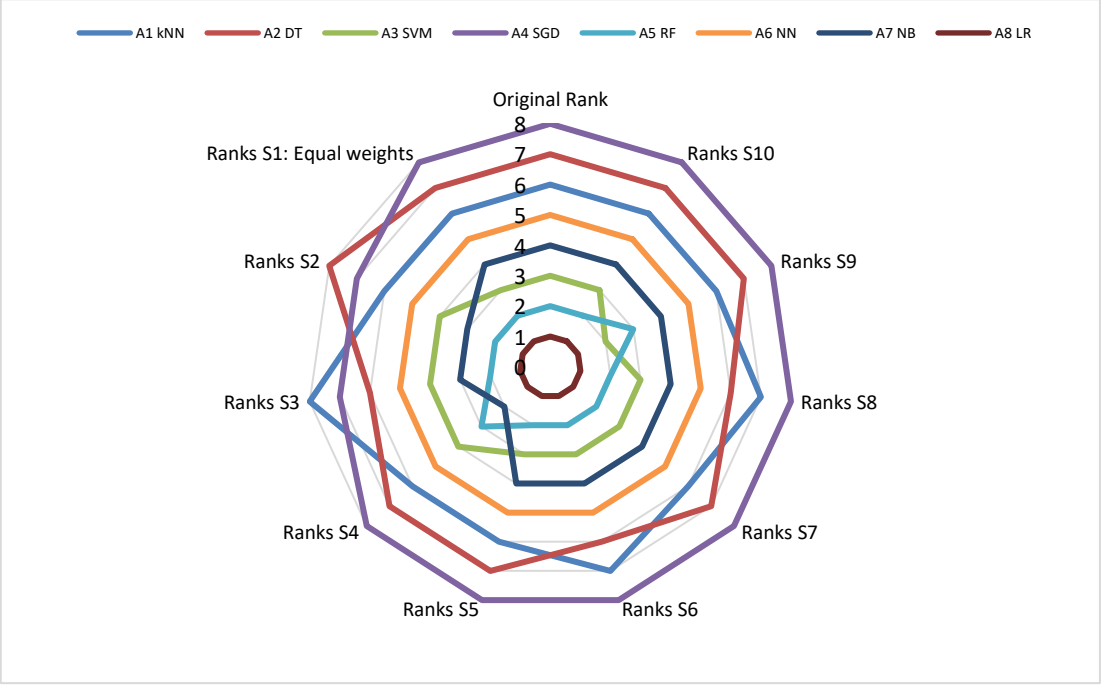

Figure 2. Rank variation under different weights for Perspective 2 (Adversarial Data).

LR maintained the top rank across all scenarios, confirming its consistent behavior under adversarial conditions. RF and SVM exhibited slight shifts, while NB moved between 2nd and 4th place, showing moderate sensitivity to criteria weighting. Models like SGD and DT consistently performed poorly.

The results of the sensitivity analysis further reinforce the central premise of this study: *Evaluating and benchmarking ML models should simultaneously consider both clean and adversarial data perspectives.* The ranking outcomes varied across the two perspectives—and in some cases, across different weighting scenarios—demonstrating that model performance is highly context-dependent. A model that ranks highest in normal conditions (e.g., SVM) may not maintain that rank under adversarial pressure (e.g., LR taking the lead). Therefore, robust and fair benchmarking requires dual-perspective evaluation frameworks, such as 2TLF-FDOSM, supplemented with sensitivity analysis to account for expert uncertainty and shifting priorities.

3.5. Explainability Analysis

This section presents an explainability analysis of the LR model trained on PCA-transformed features, evaluated both before and after applying adversarial attacks. Four complementary explainability techniques were used to provide both local and global insights into model behaviour:

1- LIME for analysing local feature contributions.
2- PFI for assessing global feature importance.
3- IG for identifying individual feature attributions.
4- PDP for visualizing how changes in features affect prediction probabilities.

This analysis focuses on the predictions generated by the LR model, which achieved the best performance among the tested models, both before and after adversarial perturbations. The model was trained using PCA-fused features to reduce dimensionality and improve interpretability. For each triage level (1, 2, and 3), predictions were analyzed before and after the adversarial attack. The applied methods were used to interpret feature behaviors and their sensitivity to adversarial changes, highlighting shifts in feature importance and prediction stability under different conditions.

3.5.1. LIME: Local Feature Explanations

LIME interprets the LR model's predictions by approximating it locally with an interpretable model using perturbed input data. It was applied to triage levels 1, 2, and 3 under two conditions: before and after adversarial attacks.

Pre-Adversarial LIME Explanations

In this running sample, the accuracy of the LR model before the adversarial attack was 0.9423. The following subsections analyze how LIME explanations were influenced:

- **Triage Level 1:** The prediction probabilities for triage level 1 are shown in Figure 3, and the PCA feature values for this sample are listed in Table 7. The LIME explanation for these features is presented in Figure 4.

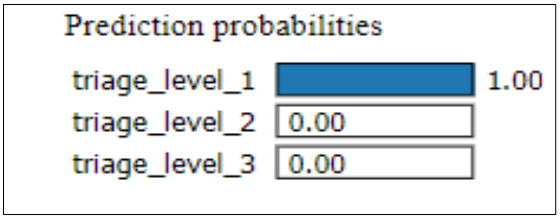

Figure 3. Prediction probabilities of the LR model for triage level 1.

TABLE VII. PCA FEATURE VALUES FOR THE CURRENT RUNNING SAMPLE OF TRIAGE LEVEL 1 BEFORE THE ATTACK

Feature	Value
PC1	4.76
PC4	-0.80
PC6	-0.16
PC12	-1.21
PC2	-0.81
PC7	-0.52
PC8	-0.75
PC10	-0.92
PC9	-0.01
PC3	0.50
PC5	-0.61
PC11	1.80

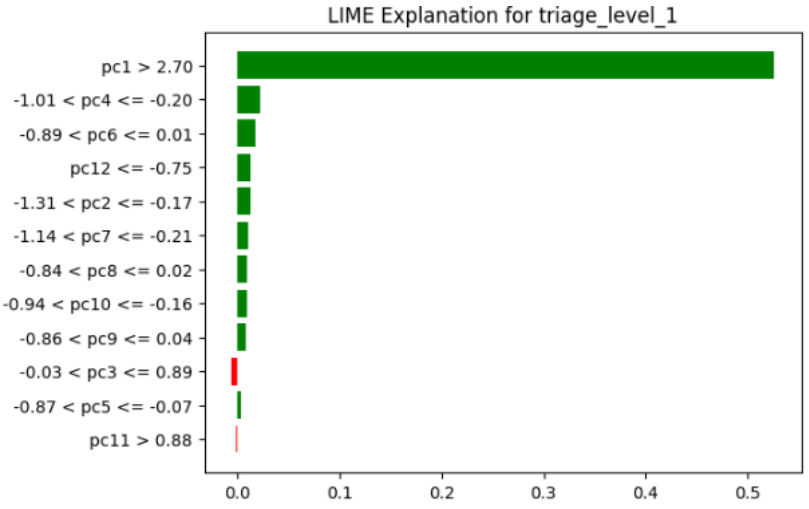

Figure 4. LIME explanation for the prediction of triage level 1 before the attack.

In Figure 4, the range of values represents the positive and negative contributions of the feature to the prediction result of triage level 1. When PC3 value is between -0.03 < PC3 <= 0.89, and PC11 value is greater than 0.88, it leads to strong negative contributions. As for the rest of the figures, PC1, PC2, PC4, PC5, PC6, PC7, PC8, PC9, PC10, and PC12, their range values in the figure show strong positive contributions when predicting triage level 1, and the highest value for positive contributions is for PC1.

- **Triage Level 2:** The prediction probabilities for triage level 2 are shown in Figure 5, and its PCA feature values are listed in Table 8. The corresponding LIME explanation is displayed in Figure 6.

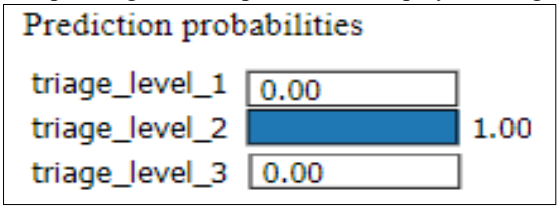

Figure 5. Prediction probabilities of the LR model for triage level 2.

TABLE VIII. PCA FEATURE VALUES FOR THE CURRENT RUNNING SAMPLE OF TRIAGE LEVEL 2 BEFORE THE ATTACK

Feature	Value
PC1	4.76
PC3	0.50
PC7	-0.52
PC6	-0.16
PC12	-1.21
PC4	-0.80
PC9	-0.01
PC8	-0.75
PC5	-0.61
PC11	1.80
PC2	-0.81
PC10	-0.92

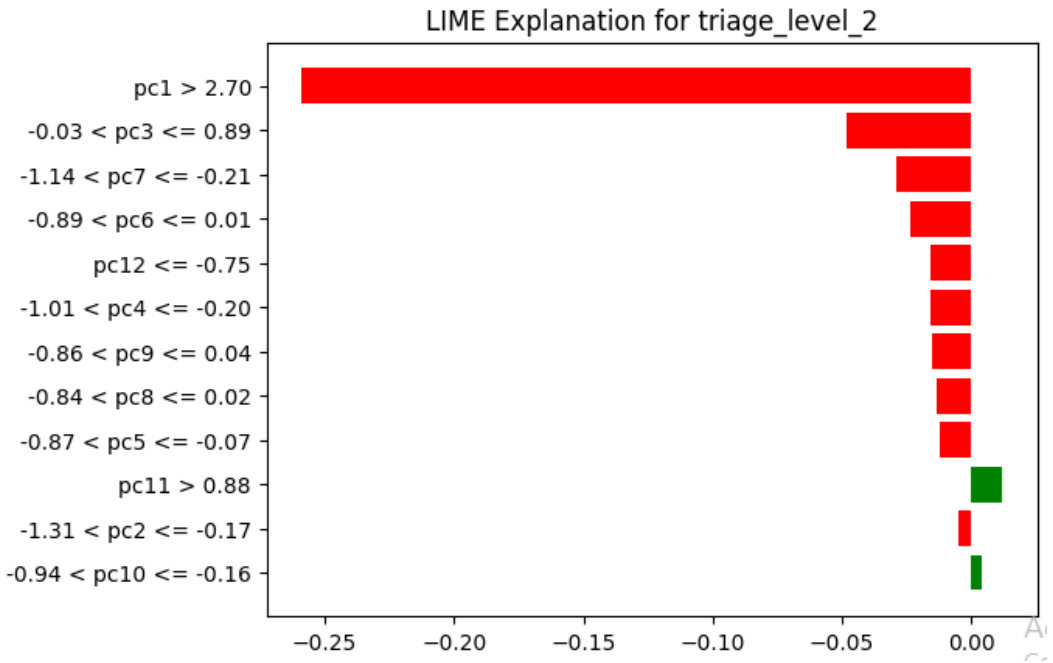

Figure 6. LIME explanation for the prediction of triage level 2 before the attack.

In Figure 6, the range of values represents the positive and negative contributions of the feature to the prediction result of triage level 2. PC11 and PC10 have a strong positive contribution. When the value of PC11 is greater than 0.88 and PC10 values are between (-0.94 < PC3 <=- 0.16). for the rest of the features, their range values in the figure show strong negative contributions when predicting triage level 2, and the highest value for negative contributions is for PC1.

- **Triage Level 3**: The prediction probabilities for triage level 3 are shown in Figure 7, and its PCA feature values are summarized in Table 9. The LIME explanation for this case is presented in Figure 8.

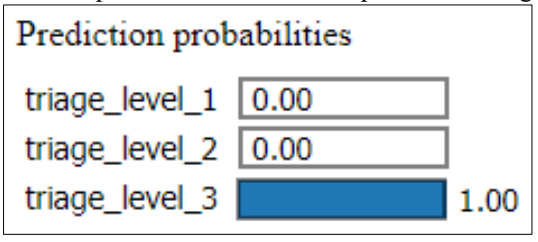

Figure 7. Prediction probabilities of the LR model for triage level 3.

TABLE IX. PCA FEATURE VALUES FOR THE CURRENT RUNNING SAMPLE OF TRIAGE LEVEL 3 BEFORE THE ATTACK

Feature	Value
PC1	4.76
PC3	0.50
PC11	1.80
PC6	-0.16
PC4	-0.80
PC5	-0.61
PC9	-0.01
PC8	-0.75
PC12	-1.21
PC7	-0.52
PC10	-0.92
PC2	-0.81

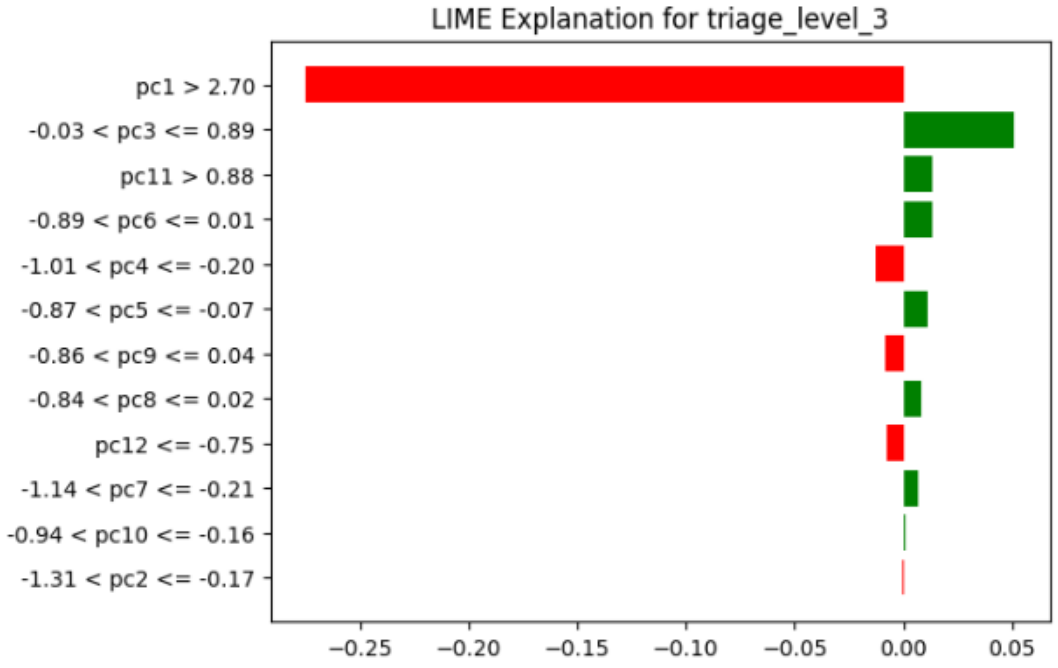

Figure 8. LIME explanation for the prediction of triage level 3 before the attack.

In the Figure 8, PC3, PC11, PC6, PC5, PC8, PC7 and PC10 have a strong positive contribution. For the rest of the features, their range values in the figure show strong negative contributions when predicting triage level 3, and the highest value for negative contributions is for PC1.

Post-Adversarial LIME Explanations

After applying a query-based adversarial attack, the LR model's accuracy declined to 0.8572. The attack perturbs the input features, potentially altering the LIME explanations. The following sections analyze the impact of the attack on LIME explanations across different triage levels.

- **Triage Level 1:** The local sample feature values for triage level 1 after the attack are presented in Table 10, and the LIME explanation is shown in Figure 9.

TABLE X. PCA FEATURE VALUES FOR THE CURRENT SAMPLE OF TRIAGE LEVEL 1 AFTER THE ADVERSARIAL ATTACK

Feature	Value
PC1	4.76
PC7	-0.52
PC6	-0.16
PC4	-0.80
PC2	-0.81
PC12	-1.21
PC10	-0.92
PC11	1.80
PC3	0.50
PC9	-0.01
PC5	-0.61
PC8	-0.75

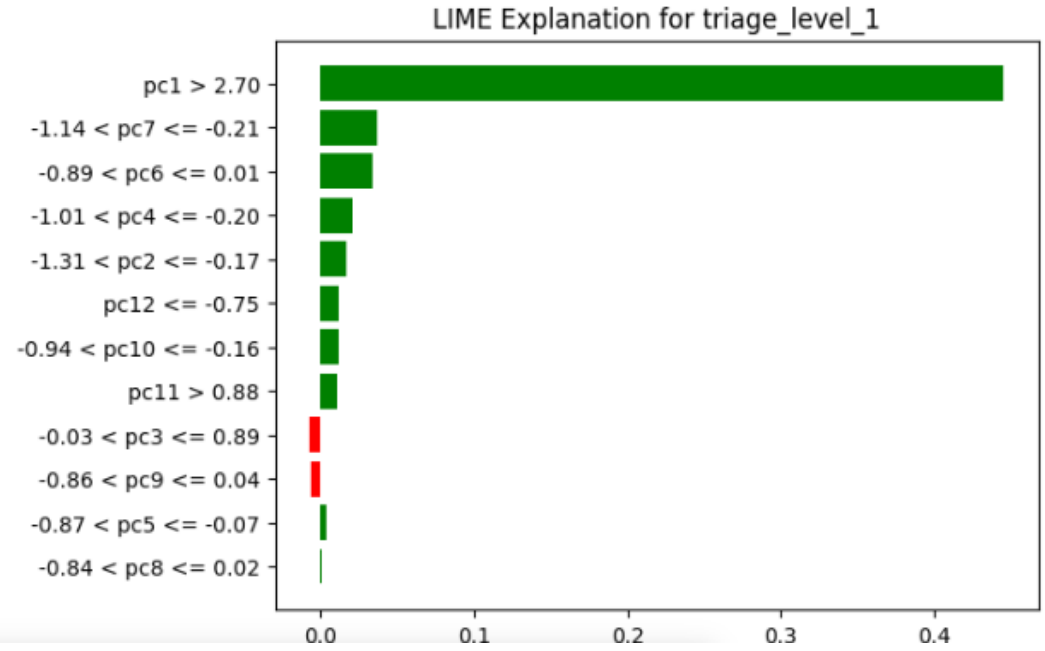

Figure 9. LIME explanation for the prediction of triage level 1 after the adversarial attack.

In the Figure 9, PC3, PC9 have a negative contribution with range values for PC3 between -0.03 < PC3 <= 0.89 and PC9 -0.86 < PC9 <= 0.04 respectively. For the rest of the features, their range values in the figure show positive contributions when predicting triage level 1, and the highest value for positive contributions is also for PC1.

- **Triage Level 2**: The feature values for triage level 2 after the attack are shown in Table 11, with the corresponding LIME explanation displayed in Figure 10.

TABLE XI. PCA FEATURE VALUES FOR THE CURRENT SAMPLE OF TRIAGE LEVEL 2 AFTER THE ADVERSARIAL ATTACK

Feature	Value
PC1	4.76
PC3	0.50
PC7	-0.52
PC5	-0.61
PC6	-0.16
PC2	-0.81
PC12	-1.21
PC11	1.80
PC4	-0.80
PC8	-0.75
PC9	-0.01
PC10	-0.92

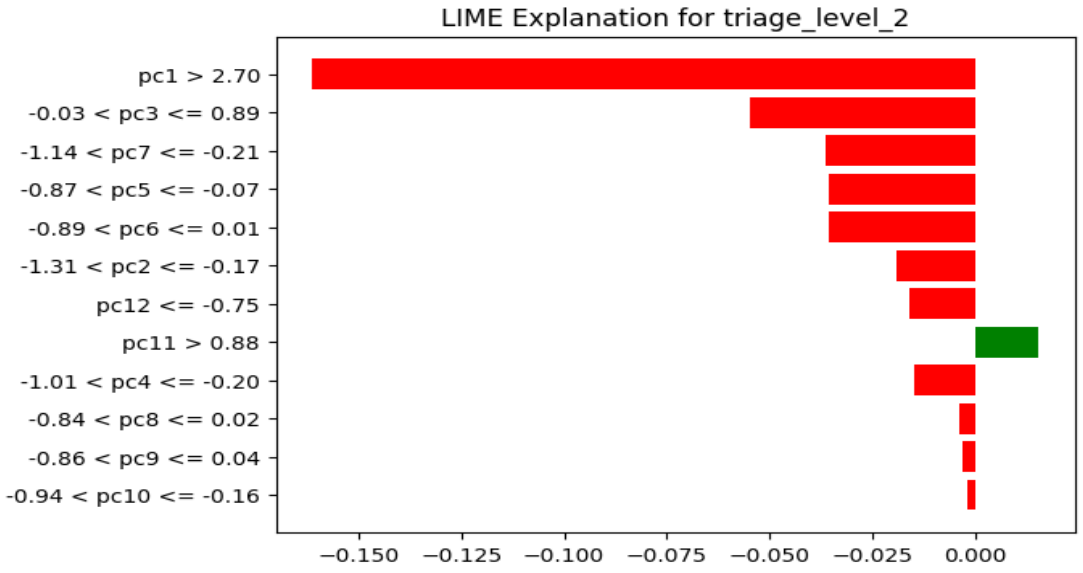

Figure 10. LIME explanation for the prediction of triage level 2 after the adversarial attack.

In Figure 10, only PC11 has a positive contribution with range values greater than 0.88. For the rest of the features, their range values in the figure show negative contributions when predicting triage level 2, and the highest value for negative contributions is also for PC1.

- **Triage Level 3**: The feature values for triage level 3 after the attack are listed in Table 12, and the LIME explanation is shown in Figure 11.

TABLE XII. PCA FEATURE VALUES FOR THE CURRENT SAMPLE OF TRIAGE LEVEL 3 AFTER THE ADVERSARIAL ATTACK

Feature	Value
PC1	4.76
PC3	0.50
PC5	-0.61
PC11	1.80
PC4	-0.80
PC2	-0.81
PC9	-0.01
PC7	-0.52
PC6	-0.16
PC8	-0.75
PC10	-0.92
PC12	-1.21

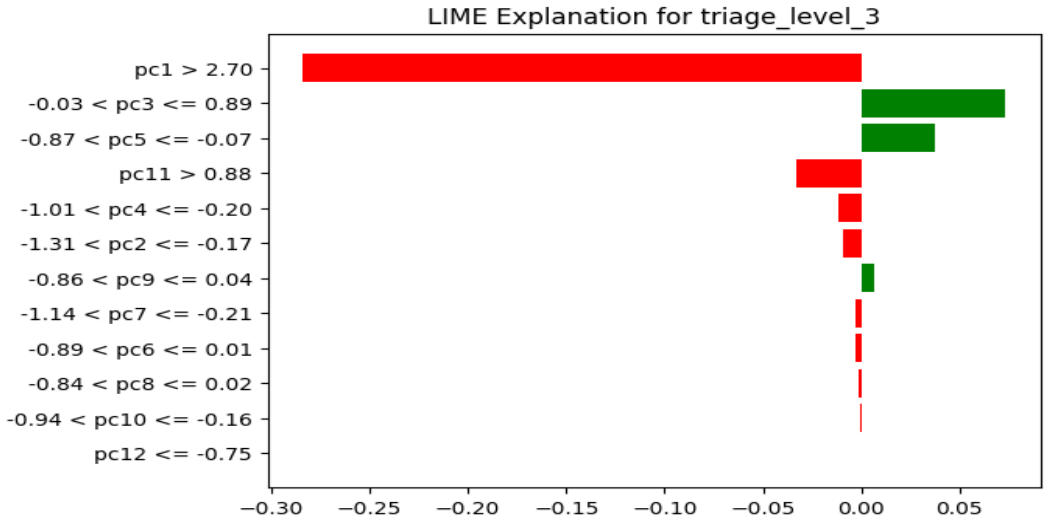

Figure 11. LIME results for the prediction of triage level 3 after an attack

In the Figure 11, PC3, PC5 and PC9 have a positive contribution with range values for PC3 between -0.03 < PC3 <= 0.89, PC5 - -0.87 < PC5 <= -0.07 and PC9 -0.86 < PC9 <= 0.04 respectively. For the rest of the features, their range values in the figure show negative contributions when predicting triage level 3, and the highest value for negative contributions is also for PC1.

In conclusion, PC1 has the strongest positive contribution in predicting triage level 1, while also making a strong negative contribution in predicting triage levels 2 and 3, even before or after applying an adversarial attack.

3.5.2 Permutation Feature Importance (PFI)

PFI offers a global perspective on feature importance by measuring the effect of randomly shuffling each feature on the model's performance. This technique is particularly informative for linear models such as LR.

Pre-Adversarial PFI Explanations

Figure 12 displays the PFI results before the adversarial attack. The plot clearly shows that PC1 is the dominant feature, exhibiting the highest importance (~0.55). In contrast, other features—such as PC5, PC12, PC2, and PC4—show minor but non-negligible positive contributions. Features such as PC10, PC9, PC3, and PC8 exhibit zero contribution to the model's predictive performance.

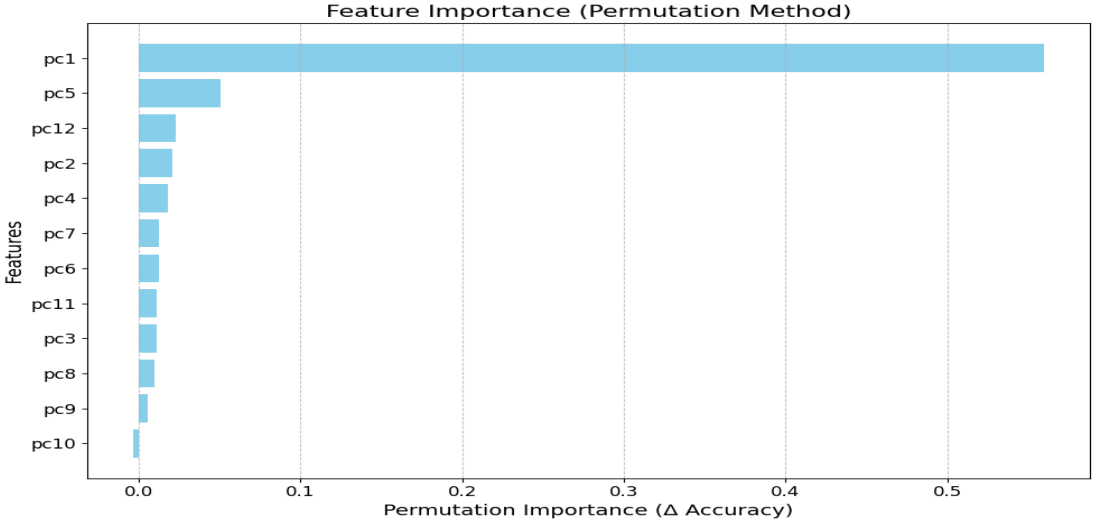

Figure 12. PFI before the adversarial attack.

Post-Adversarial PFI Explanations

As shown in Figure 13, after the adversarial attack, PC1 continues to dominate with the highest importance (~0.52), although its value slightly decreases. However, the attack significantly affects weaker features. Features such as PC2, PC5, and PC12 show a notable drop in importance. Some features, including PC6, PC8, and PC11, even exhibit negative permutation importance—indicating that shuffling these features actually improved the model's performance, which is a sign of instability or overfitting. Features PC3, PC10, and PC9 remain near zero, similar to the pre-attack state.

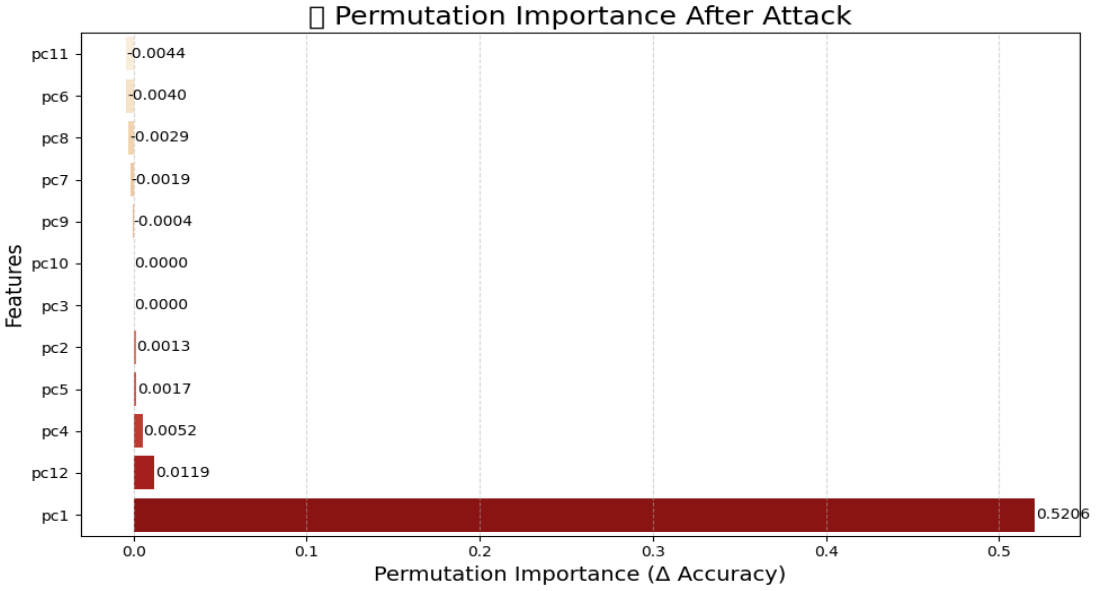

Figure 13. PFI after the adversarial attack.

The results suggest that the LR model demonstrates partial robustness to adversarial attacks, particularly regarding its most predictive feature (PC1). While PC1 remains stable and consistently important before and after the attack, the adversarial perturbations primarily destabilize weaker features. These findings reveal two key insights: (1) The LR model maintains stable weights for its most influential feature, making it relatively robust in this respect. (2) Adversarial attacks disproportionately affect less important features, reducing their importance further or even inverting their impact, as shown by the negative PFI scores.
This highlights the vulnerability of weaker components within the model while reinforcing the stability of the core predictive feature (PC1).

3.5.3 Integrated Gradients (IG)

IG is the third explainability method used in this analysis. It helps quantify the contribution of each input feature to a model's prediction by attributing changes in the output to the features along a baseline-to-input path.

Pre-Adversarial IG Explanations

Figures 15, 16, and 17 illustrate the IG results for the three triage levels before adversarial attacks.

- **Triage Level 1** (Figure 15): The dominant features, ranked by IG value, are PC1, PC5, PC12, PC4, PC2, and PC6. These features have the highest attributions, indicating their strong influence on the model's predictions for this class.

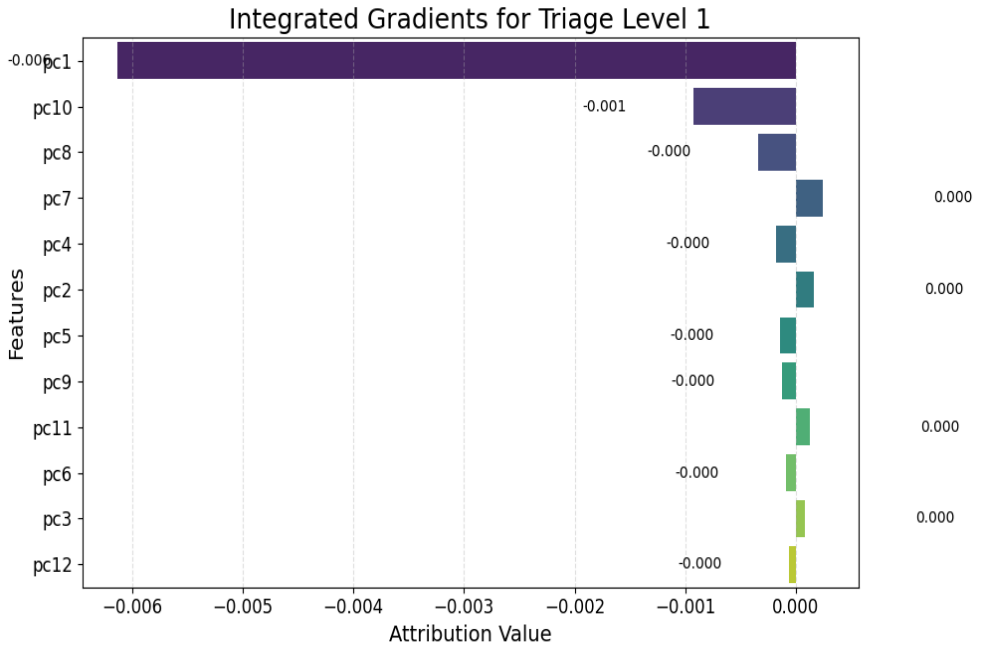

Figure 15. IG for triage level 1 before the adversarial attack.

- **Triage Level 2** (Figure 16): The most influential features are PC1, PC2, PC5, PC6, PC4, and PC9. Among these, PC1 shows the strongest contribution (either positive or negative), followed by PC2, with the remaining features contributing moderately.

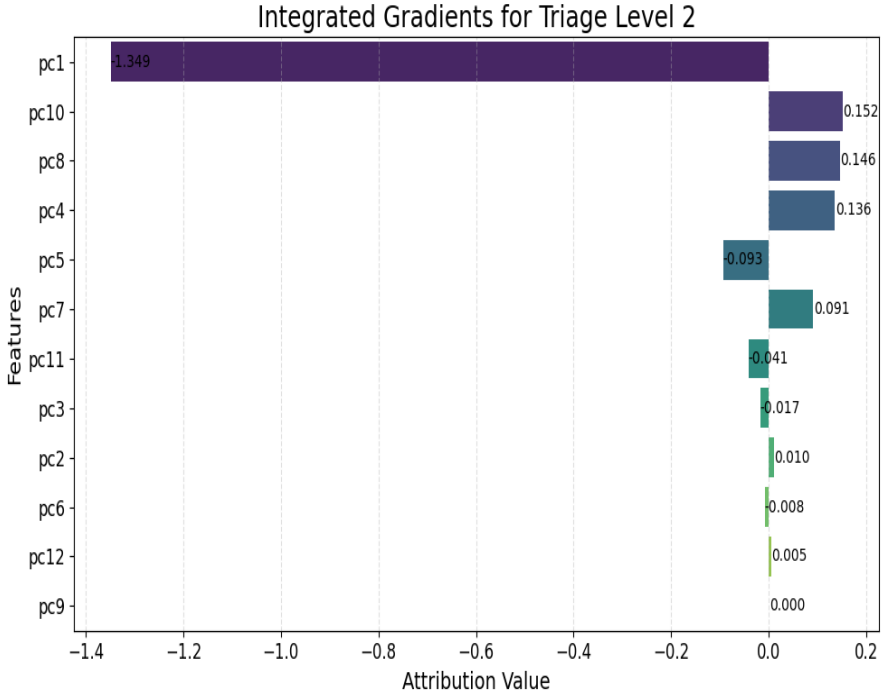

Figure 16. IG for triage level 2 before the adversarial attack.

- **Triage Level 3** (Figure 17): Top features include PC1, PC5, PC12, PC3, PC4, and PC6. Once again, PC1 remains dominant, with other features sharing moderate importance.

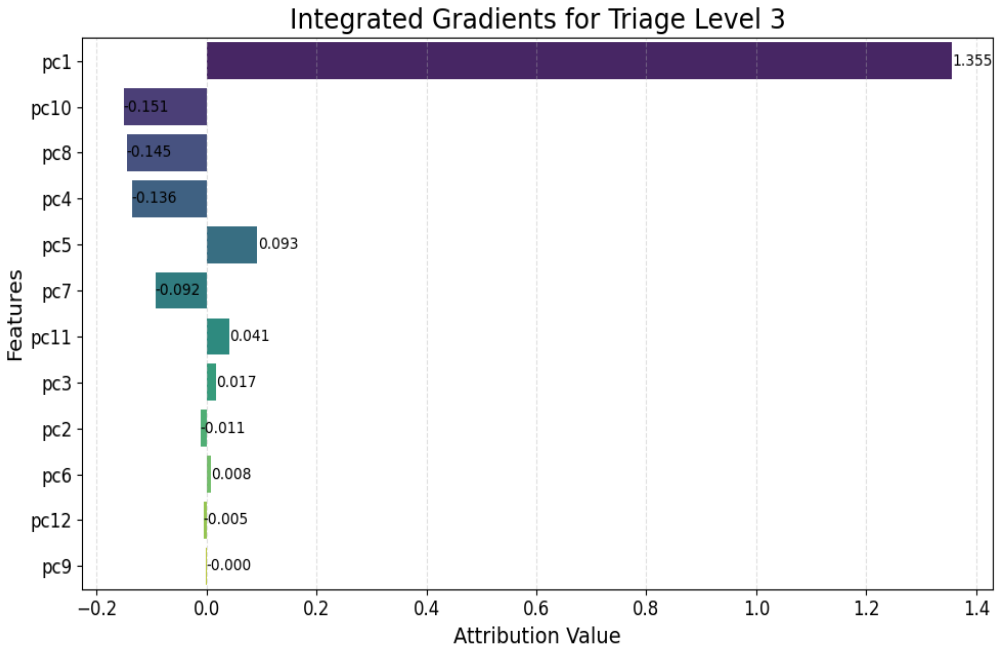

Figure 17. IG for triage level 3 before the adversarial attack.

Pre-Adversarial IG Explanations

Figures 18, 19, and 20 present the IG results for the three triage levels after adversarial attacks.

- **Triage Level 1** (Figure 18):

The ranking of feature importance shifts slightly. The key features become PC1, PC5, PC4, PC12, PC2, and PC8. PC1 remains dominant, but PC8 shows a slight increase in importance compared to before the attack.

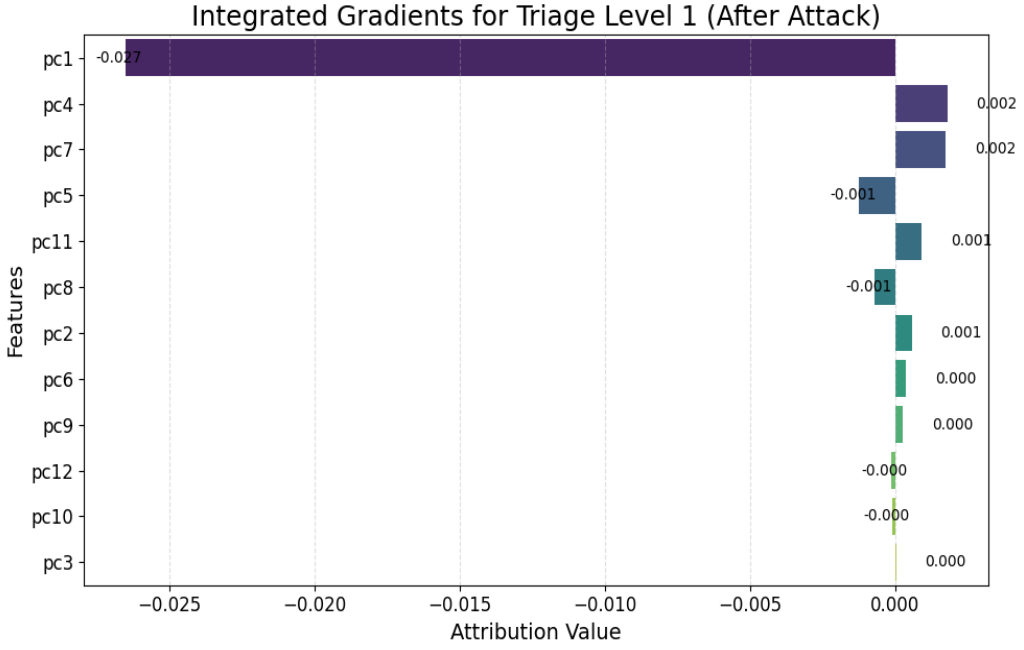

Figure 18. IG for triage level 1 after the adversarial attack.

- **Triage Level 2** (Figure 19): Feature attributions shift more noticeably. The top features now include PC1, PC5, PC4, PC6, PC2, and PC3. The model appears to rely more on PC4 and PC5 post-attack, while the importance of

PC2 diminishes. This suggests that although the model may still yield correct predictions, its internal decision-making becomes less stable under attack.

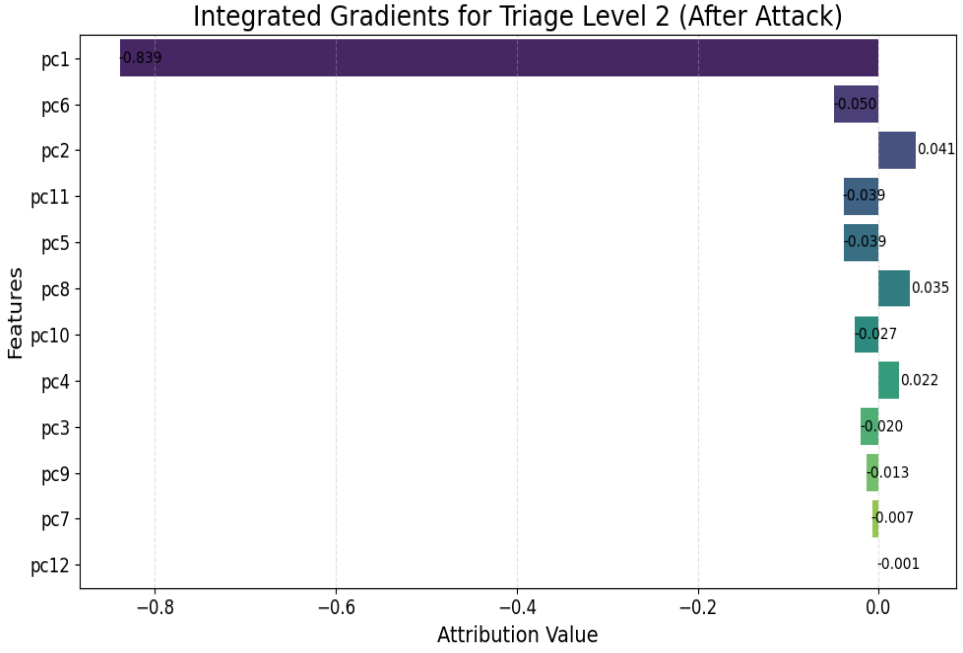

Figure 19. IG for triage level 2 after the adversarial attack.

- **Triage Level 3** (Figure 20): Some attribution shifts are observed, with importance shifting from PC3 to PC4, while PC1 remains the dominant feature.

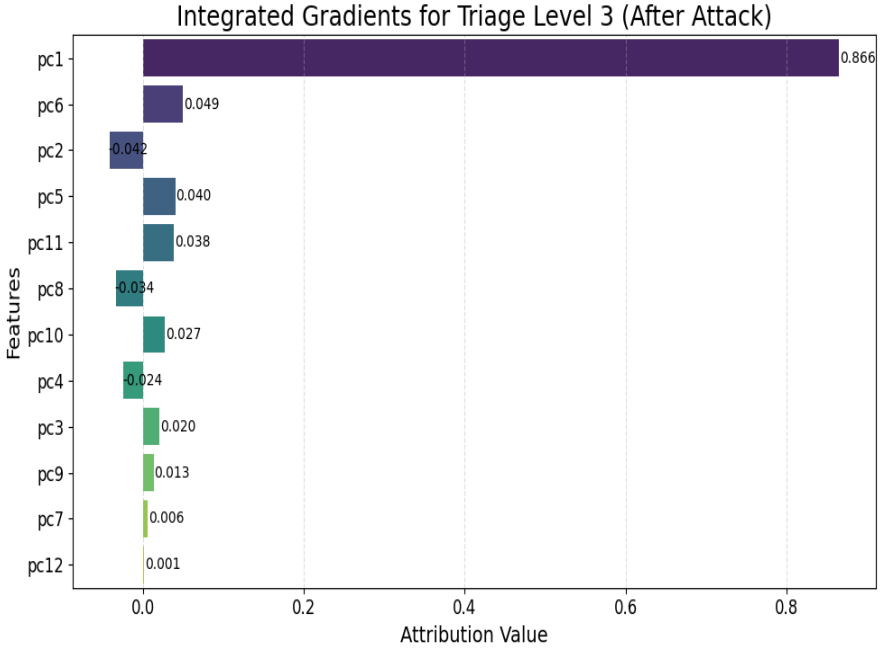

Figure 20. IG for triage level 3 after the adversarial attack.

Across all triage levels, PC1 consistently remains the most influential feature, both before and after the adversarial attacks. However, adversarial perturbations lead to notable shifts in how the model attributes importance to the weaker features. Features such as PC5, PC4, PC2, and PC3 show varying degrees of influence after attacks. The adversarial attack causes the model to redistribute its reliance on these weaker features, signaling its susceptibility to small input

changes. This analysis highlights that while core features, such as PC1, remain stable, the model's dependency on secondary features is more fragile under adversarial conditions.

3.5.4 Partial Dependence Plot (PDP)

PDP provide a global perspective on how variations in individual features (PC1 to PC12) affect the model's predicted probability for each triage level. This method is particularly useful for visualizing feature impacts both before and after adversarial attacks.

Pre-Adversarial PDP Explanations

- **Triage Level 1** (Figure 21): Before the attack, the model exhibits balanced sensitivity to multiple features. Specifically, PC2, PC3, and PC6 show moderate influence on the predicted probability for triage level 1. Their PDP curves demonstrate smooth and predictable relationships, meaning the model reacts consistently to changes in these features. Features like PC8 to PC12 show almost flat curves, indicating minimal contribution under normal conditions. This suggests that the model makes robust and well-distributed predictions for this triage level prior to adversarial perturbation.

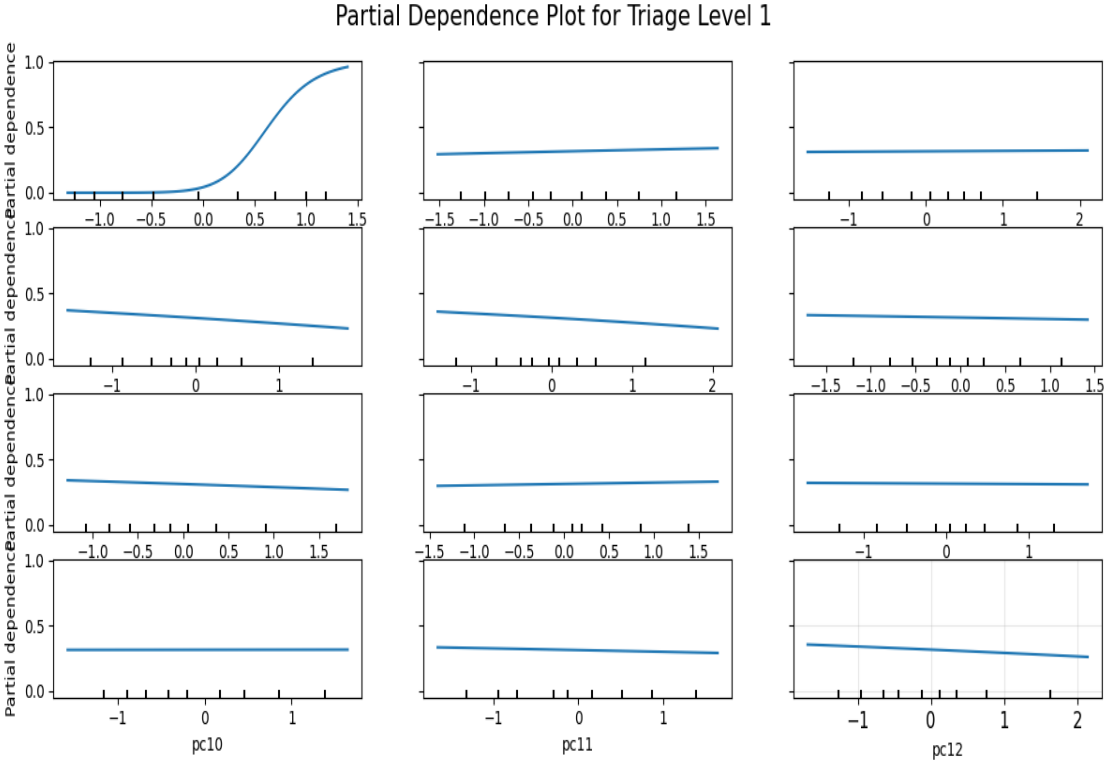

Figure 21. PDP for triage level 1 before the adversarial attack.

- **Triage Level 2** (Figure 22): Here, the model demonstrates reliance on features such as PC5, PC7, PC8, and PC9. These features exhibit non-linear yet smooth relationships, indicating that the model is moderately sensitive to specific value ranges for these features when predicting triage level 2. In contrast, features like PC1, PC2, and PC3 show flatter PDP curves, reflecting their relatively smaller role in this classification task.

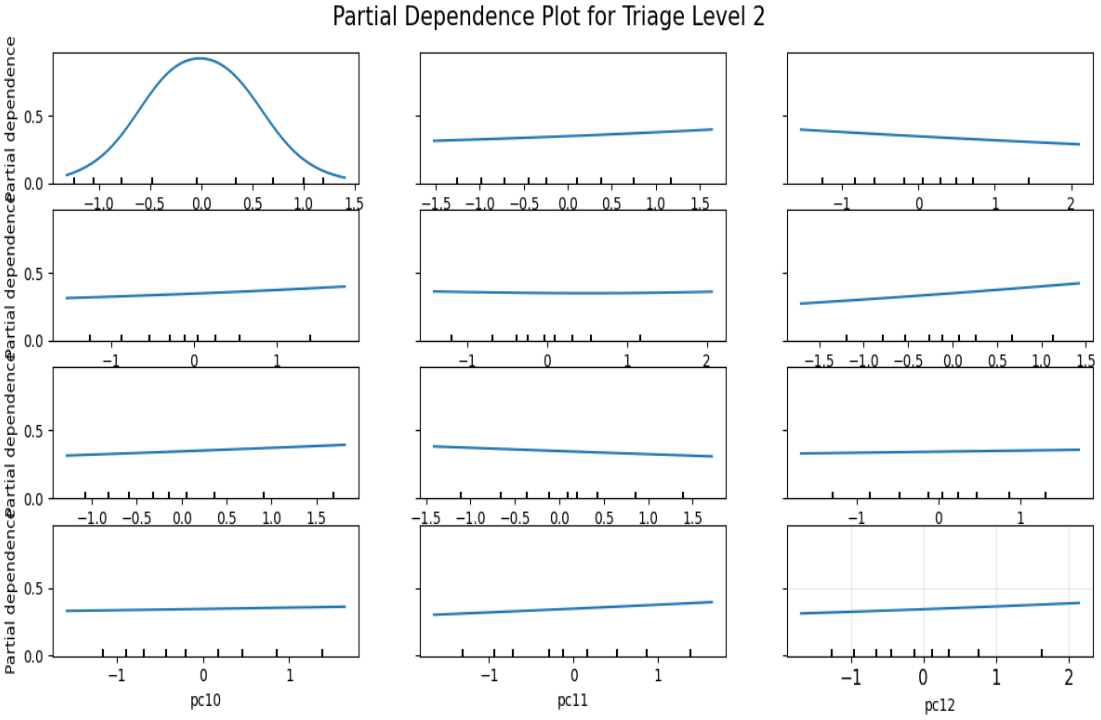

Figure 22**.** PDP for triage level 2 before the adversarial attack.

- **Triage Level 3** (Figure 23): For triage level 3, the model shows stronger dependence on higher-order principal components such as PC1. The PDP curves for these features demonstrate clear slopes, where changes in feature values significantly impact predicted probabilities. In contrast, features like PC5, PC10, PC11, and PC12 contribute minimally, as shown by their flatter PDP curves. This suggests the model effectively leverages different feature sets across triage levels for accurate classification.

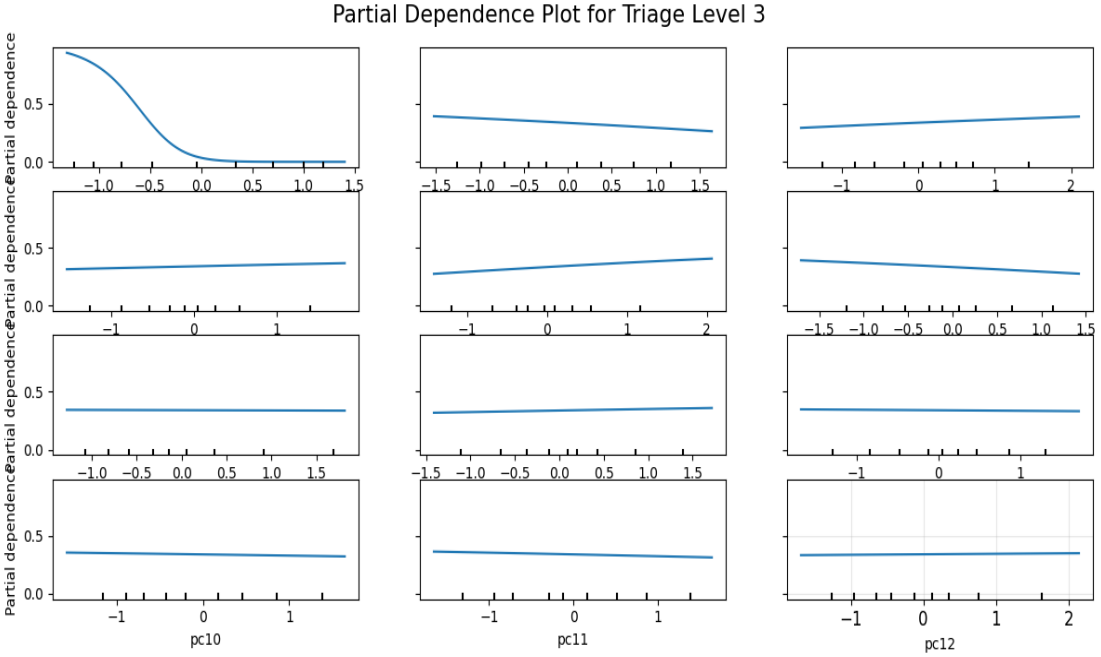

Figure 23. PDP for triage level 3 before the adversarial attack.

Post-Adversarial PDP Explanations

- **Triage Level 1** (Figure 24): After the attack, the PDP curves for PC2, PC3, and PC6 flatten noticeably, indicating reduced sensitivity to these features. Conversely, PC4 and PC5 show sharper transitions, revealing a shift in the model's attention towards fewer features under adversarial conditions. This reflects a narrowing of the model's focus and an over-reliance on select features following the attack.

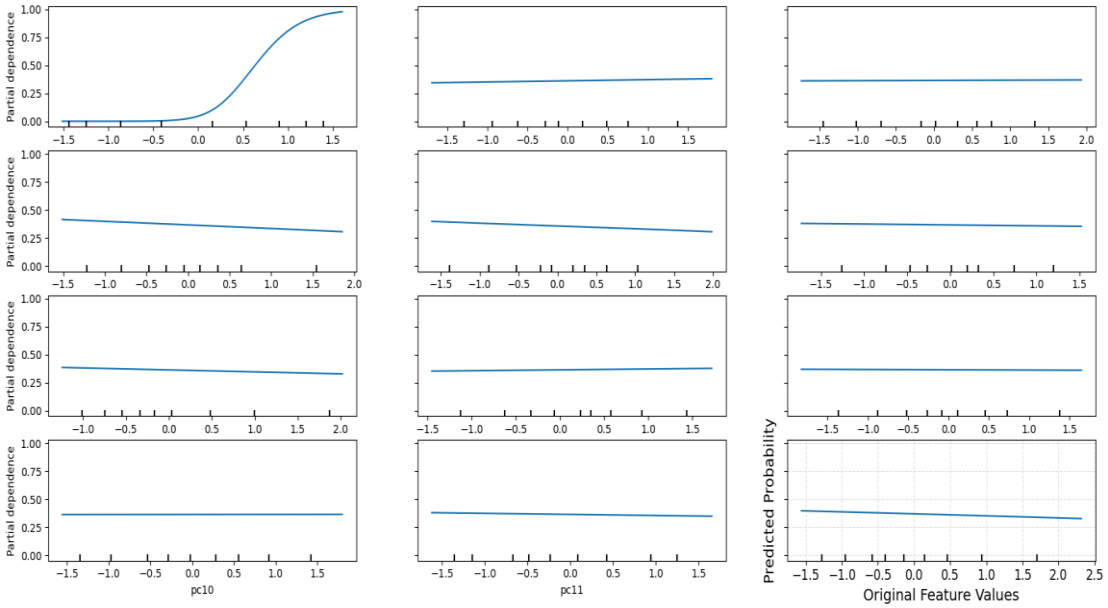

Figure 24. PDP for triage level 1 after the adversarial attack.

- **Triage Level 2** (Figure 25): The PDP curves now exhibit irregular patterns, particularly for PC7, PC8, and PC9, where even minor feature changes lead to significant shifts in predicted probability. This indicates that adversarial perturbation has made the model's decision boundary for triage level 2 highly unstable and vulnerable to noise.

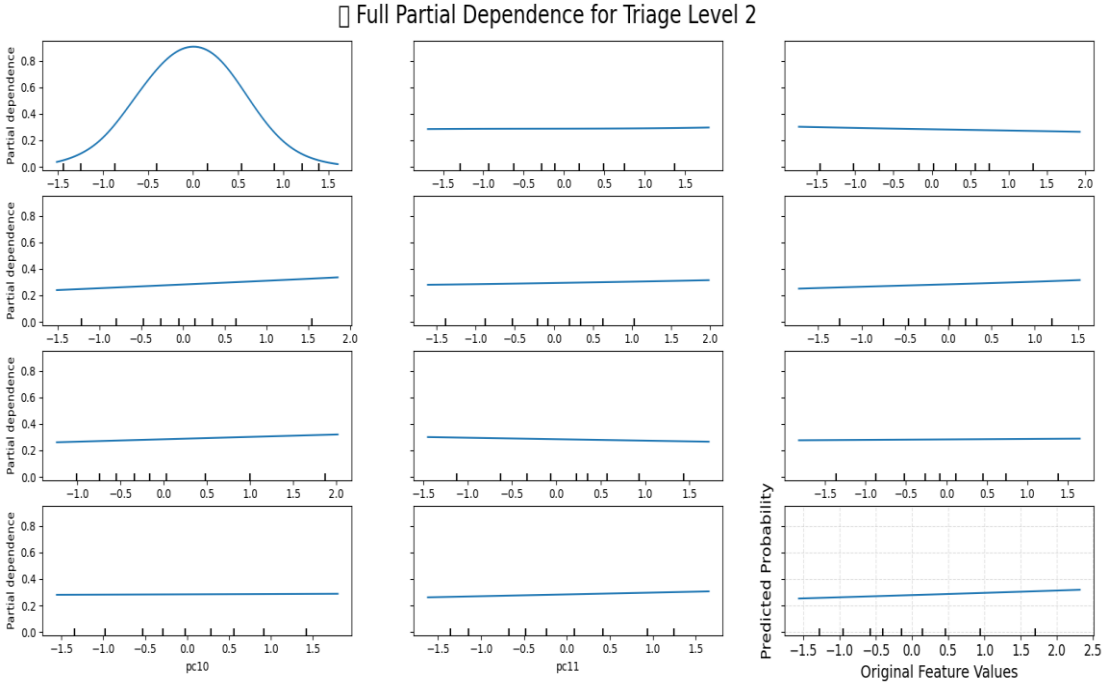

Figure 25**.** PDP for triage level 2 after the adversarial attack.

- **Triage Level 3** (Figure 26): For triage level 3, the model shows amplified sensitivity to features like PC1 and PC2, where extreme feature values lead to large fluctuations in predictions. Meanwhile, features such as PC10, PC11, and PC12 remain largely flat, indicating diminished importance. This highlights increased model fragility and reduced generalizability for triage level 3 after the attack.

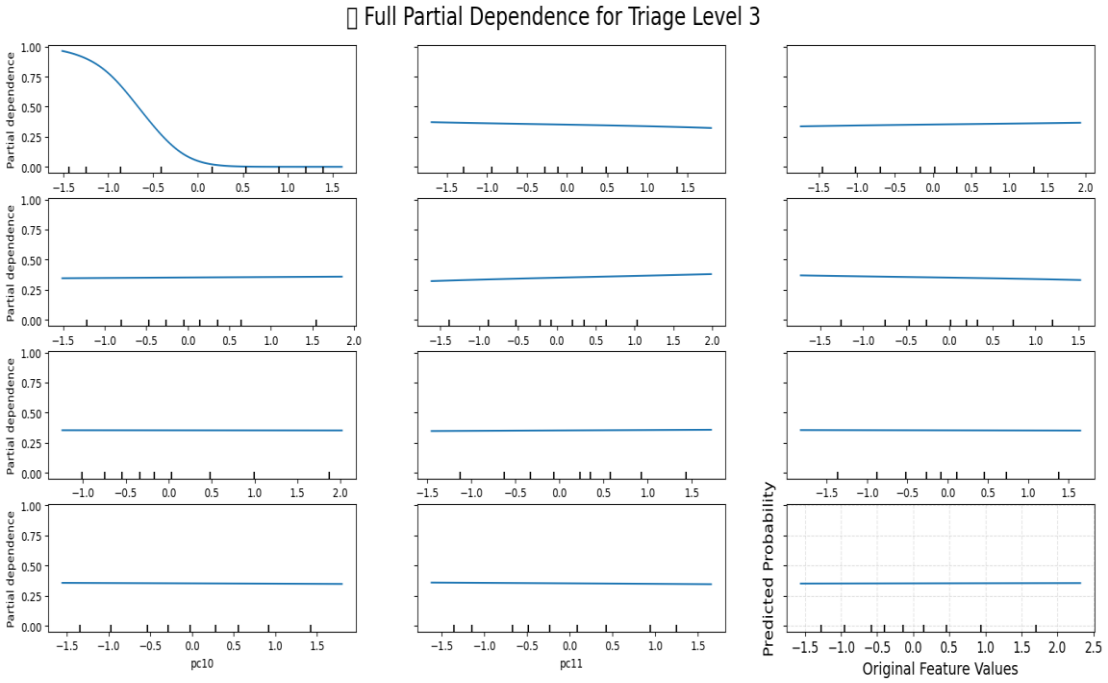

Figure 26. PDP for triage level 3 after the adversarial attack.

The adversarial attack substantially alters the feature importance landscape across all triage levels. Before the attack, the model exhibited stable and well-generalized decision boundaries, characterized by distributed feature reliance. After the attack, this balance is disrupted. The model becomes overly dependent on a smaller number of features, exhibits increased sensitivity to minor changes in features, and demonstrates erratic behavior.
This highlights how adversarial attacks can compromise not only prediction accuracy but also the interpretability, stability, and trustworthiness of ML models—particularly in critical applications, such as triage level prediction for autism-related cases.

4. STATE OF THE ART: COMPARISON OF STUDIES

In this section, a detailed comparison of the proposed framework with existing literature is carried out through a checklist benchmarking method. One of the current methodologies used frequently in the literature is checklist benchmarking. This technique involves assessing various significant checklists presented as factors to highlight the originality of the proposed work. In these contexts, this section presents a structured classification of the comparison points used to evaluate the proposed decision-making framework. The points are grouped into Theoretical-Level and Application-Level to distinguish between methodological innovations and practical implementations. The definitions of these checklists are provided below, and Table 5.13 illustrates how the proposed framework contributes to the existing body of literature based on the results obtained.

- **Theoretical-Level Comparison Points**

1. **Development of MCDM Selection Method:** This element highlights the difference between employing existing methods and creating new methods for choosing the best machine learning model for ASD triage. Model selection is a key challenge tackled in the current research, advocating for the application of an appropriate MCDM method to successfully address this concern. Thus, research that focuses on developing new MCDM methods is deemed to meet this comparison criterion.
2. **Decision Matrix Development:** Building on the context of "Selection Method Development," this aspect relates to the formulation of new decision matrices or refinements to this essential component of the decision-making process. It stresses the importance of innovating or enhancing decision matrices, which are vital to the research methodology.

3. **Sensitivity Analysis:** Assesses how changes in criteria weights affect the ranking outcomes, evaluating the robustness and stability of the decision-making process.
4. **Multi-View Decision Modelling:** A theoretical construct addressing complex decision scenarios involving multiple conflicting viewpoints and diverse evaluation and benchmarking. It includes:
    - **Criteria Weighted Handling:** Uses fuzzy and linguistic weighting mechanisms to reconcile varying expert opinions on criteria importance.
    - **Criteria Trade-offs Handling:** Models the interplay between mutually influencing or competing criteria (e.g., Accuracy vs. training time with consideration of adversarial and normal).
    - **Criteria Conflicts Handling:** Resolves criteria-level contradictions through compromise solutions and fuzzy consensus mechanisms.

- **Application-Level Comparison Points**

5. **Normal vs. Adversarial Perspectives**: This aspect indicates how the proposed work accentuates the importance of considering both normal test instances and adversarial attack examples while developing ML models for the triage of autistic patients. It compares model performance under typical conditions and in the face of adversarial input variations, demonstrating model robustness.
6. **Fusion Improvements**: This aspect highlights the importance of feature fusion in improving the creation of machine learning models. Therefore, it emphasizes the methodology used for preprocessing the ASD dataset, indicating that studies addressing this checklist comparison point should consider feature fusion during the pre-processing phase.
7. **Integrated Medical and Sociodemographic Features:** The integration of both medical and sociodemographic features has shown its relevance in the detection, diagnosis, and triage of ASD patients. Consequently, this aspect underscores the need for integrating both types of features when developing the study framework for evaluating autistic patients.
8. **Explainability Analysis:** Uses explainable AI tools to interpret model outcomes, improving decision transparency and clinician trust.
9. **Individual and Group Ranks:** Analyzes rankings from individual experts versus group consensus, emphasizing collaborative decision-making.
10. **Use of Real-World Dataset:** The methodology is validated using actual ASD datasets, which include medical and Sociodemographic features, to demonstrate its real-world applicability and relevance.

TABLE XIII. COMPARISON OF PERSPECTIVES AND POINTS IN THE BENCHMARKS AND PROPOSED FRAMEWORK

Comparison Points/Benchmarks				**Benchmark#1** [13]	**Benchmark #2** [14]	**Benchmark #3** [4]	**Benchmark #4** [15]	**Benchmark #5** [16]	**Proposed Framework**
Theoretical-Level	**1**	**Development of MCDM Selection Method**		✓	χ	χ	χ	χ	✓
	2	**Decision Matrix Development**		χ	χ	χ	χ	χ	✓
	3	**Sensitivity Analysis**		✓	χ	χ	χ	χ	✓
	4	**Multi-View Decision Modeling**	Criteria Weighted Handling	✓	χ	χ	χ	χ	✓
			Criteria Trade-offs Handling	✓	χ	χ	χ	χ	✓
			Criteria Conflicts Handling	✓	χ	χ	χ	χ	✓
Application-Level	**5**	**Normal/Adversarial Perspectives**	a) Normal Test Examples	✓	✓	✓	✓	✓	✓
			b) Adversarial Attack Examples	χ	χ	χ	χ	χ	✓

	6	**Fusion Improvements**		χ	χ	χ	χ	χ	✓
	7	**Integrated Medical and Sociodemographic Features**	a) Medical Features	✓	χ	χ	χ	✓	✓
			b) Sociodemographic Features	✓	✓	✓	✓	✓	✓
	8	**Explainability Analysis**		χ	χ	χ	χ	✓	✓
	9	**Individual and Group Ranks**		✓	χ	χ	χ	χ	✓
	10	**Use of Real-World ASD Medical and Sociodemographic Dataset**		✓	χ	χ	χ	✓	✓
Total score				**71.42%**	**14.28%**	**14.28%**	**14.28%**	**35.71%**	**100%**
Difference				**28.58%**	**85.72%**	**85.72%**	**85.72%**	**64.29%**	

This comparison table clearly positions the proposed framework as a comprehensive and balanced solution, integrating both theoretical innovations and practical enhancements. The analysis is structured into two major layers: Theoretical Level and Application Level, to systematically evaluate advancements over five recent and relevant benchmarks. Only the proposed framework introduces a two-tier layer fuzzy decision-optimised selection model (2TLFFDOSM), enabling a structured and optimised decision-making process. All existing benchmarks rely on conventional or single-layer MCDM strategies without advanced selection optimization. The proposed approach is unique in explicitly developing a customised decision matrix tailored to handle both normal and adversarial perspectives across medical and sociodemographic dimensions. This ensures contextual relevance and enhances decision accuracy.
Despite its importance, sensitivity analysis is often overlooked by most benchmarks, except for one (Benchmark #1). The proposed framework incorporates a full-scale sensitivity analysis to examine the robustness of decision rankings under varying weights and perturbations.

A critical and novel theoretical contribution is the inclusion of Multi-View Decision Modelling, broken into three distinct but related subdimensions: criteria-weighted conflict handling to address weight-based discrepancies between conflicting criteria. criteria trade-offs handling to balance trade-offs among features such as accuracy vs. fairness or complexity vs. interpretability. Handling criteria conflicts to resolve direct contradictions among criterion values, especially relevant in adversarial contexts. This structured decomposition adds depth to the theoretical underpinnings, highlighting the framework's adaptability and precision in complex decision scenarios.

While all benchmarks evaluate normal data, none incorporate adversarial test cases, which are increasingly relevant in real-world AI/ML deployments, especially in healthcare. The proposed framework addresses this challenge uniquely, reinforcing its robustness against attacks and providing enhanced reliability. The framework introduces fusion-based performance improvements that integrate classifier outputs with decision modeling results. This ensures improved accuracy and alignment between prediction and prioritization, which is absent in all comparative studies. Only a few benchmarks consider both medical and sociodemographic attributes simultaneously. The proposed method systematically fuses these dimensions, capturing comprehensive patient profiles, which is crucial for holistic and equitable prioritization in ASD diagnosis and care. Despite growing emphasis on explainable AI, most benchmarks neglect it. The proposed framework incorporates both global and local explainability mechanisms to ensure transparency, particularly in the context of adversarial threat models. By providing individual and group prioritization scores, the proposed approach facilitates flexible and hierarchical decision-making, which is suitable for both personalised diagnosis and group-level policy planning. Real-world data use is a core strength of the framework. Unlike many benchmarks that use synthetic or limited data, this approach utilises a real-world ASD dataset that combines electronic medical records and sociodemographic survey data, ensuring high external validity. The results demonstrate that the proposed framework significantly outperforms existing benchmarks by offering a comprehensive, robust, and explainable decision-making solution under both normal and adversarial conditions.

5. CONCLUSION

This study evaluated the proposed decision-making framework through a series of structured analyses. The systematic ranking is validated by the 2TLFFDOSM-generated rankings, which are clustered into four performance groups based on normalized metrics using ML models. Group-wise mean ± STD values confirmed that models in G1 had the best performance, followed by G2, while G3 and G4 showed weaker results. The consistent increase in group means (G1 < G2 < G3 < G4) aligns with the generated rankings and reinforces the reliability and robustness of the proposed evaluation framework. The evaluation is presented in normal and then adversarial form, with in-depth ranking analysis of eight ML models under both normal (clean) and adversarial input conditions using the 2TLF-FDOSM. The results highlighted that model performance varies significantly between the two perspectives, thereby reinforcing the central study that robust evaluation requires considering both environments simultaneously.

A sensitivity analysis was conducted by varying the weights of nine evaluation criteria across ten scenarios. The results demonstrated the resilience and consistency of the model rankings, particularly for top-ranked models such as SVM and LR, validating the reliability of the decision-making framework even under uncertain or perturbed conditions. The explainability of the best-performing model, LR, was investigated using comprehensive AXI methods: LIME, PFI, IG, and PDP, taken into consideration pre- and post-adversarial explanations. The analysis provided insights into feature importance and decision reasoning, ensuring that the model's outputs are interpretable and trustworthy, especially when deployed in high-stakes medical environments. Finally, this study benchmarked the proposed framework against five recent studies using a multi-point checklist across theoretical and application dimensions. The proposed framework achieved 100% coverage, particularly excelling in areas such as adversarial testing, multi-view decision modeling, sensitivity analysis, and integrated feature usage. Overall, the study validated the methodological soundness and practical relevance of the proposed framework, supporting its potential for reliable and explainable ML-based triage systems in complex, real-world medical environments.

Conflicts of Interest

The authors declare no conflicts of interest.

Funding

This research received no external funding.

Acknowledgement

None

References

[1] A. C. M. do Rêgo and I. Araújo-Filho, “Leveraging artificial intelligence to enhance the quality of life for patients with autism spectrum disorder: A comprehensive review,” *Eur. J. Clin. Med.*, vol. 5, no. 5, pp. 28–38, 2024.

[2] S. S. Joudar *et al.*, “Artificial intelligence-based approaches for improving the diagnosis, triage, and prioritization of autism spectrum disorder: a systematic review of current trends and open issues,” *Artif. Intell. Rev. 2023*, pp. 1–65, Jun. 2023, doi: 10.1007/S10462-023-10536-X.

[3] G. G. Shayea *et al.*, “Fuzzy Evaluation and Benchmarking Framework for Robust Machine Learning Model in Real-Time Autism Triage Applications,” *Int. J. Comput. Intell. Syst.*, vol. 17, no. 1, p. 151, Jun. 2024, doi: 10.1007/s44196-024-00543-3.

[4] R. A. Rasul, P. Saha, D. Bala, S. M. R. U. Karim, M. I. Abdullah, and B. Saha, “An evaluation of machine learning approaches for early diagnosis of autism spectrum disorder,” *Healthc. Anal.*, vol. 5, p. 100293, 2024, doi: 10.1016/j.health.2023.100293.

[5] L. Jin, H. Cui, P. Zhang, and C. Cai, “Early diagnostic value of home video–based machine learning in autism spectrum disorder: a meta-analysis,” *Eur. J. Pediatr.*, vol. 184, no. 1, p. 37, 2024.

[6] O. S. Albahri *et al.*, “Multidimensional benchmarking of the active queue management methods of network congestion control based on extension of fuzzy decision by opinion score method,” *Int. J. Intell. Syst.*, vol. 36, no. 2, pp. 796–831, Feb. 2021, doi: 10.1002/int.22322.

[7] N. Kalid *et al.*, “Based on Real Time Remote Health Monitoring Systems: A New Approach for Prioritization ‘Large Scales Data’ Patients with Chronic Heart Diseases Using Body Sensors and Communication Technology,” *J. Med. Syst.*, vol. 42, no. 4, 2018, doi: 10.1007/s10916-018-0916-7.

[8] S. Dožić and M. Kalić, "Comparison of two MCDM methodologies in aircraft type selection problem," *Transp. Res. Procedia*, vol. 10, pp. 910–919, 2015, doi: 10.1016/j.trpro.2015.09.044.
[9] S. Moradi, H. Yousefi, Y. Noorollahi, and D. Rosso, "Multi-criteria decision support system for wind farm site selection and sensitivity analysis: Case study of Alborz Province, Iran," *Energy Strateg. Rev.*, vol. 29, p. 100478, 2020, doi: 10.1016/j.esr.2020.100478.
[10] B. Yagmahan and H. Yılmaz, "An integrated ranking approach based on group multi-criteria decision making and sensitivity analysis to evaluate charging stations under sustainability," *Environ. Dev. Sustain.*, vol. 25, no. 1, pp. 96–121, 2023, doi: 10.1007/s10668-021-02044-1.
[11] K. C. Pai, S. A. Su, M. C. Chan, C. L. Wu, and W. C. Chao, "Explainable machine learning approach to predict extubation in critically ill ventilated patients: a retrospective study in central Taiwan," *BMC Anesthesiol.*, vol. 22, no. 1, p. 75, 2022, doi: 10.1186/s12871-022-01888-y.
[12] G. Visani, E. Bagli, F. Chesani, A. Poluzzi, and D. Capuzzo, "Statistical stability indices for LIME: Obtaining reliable explanations for machine learning models," *J. Oper. Res. Soc.*, vol. 73, no. 1, pp. 91–101, 2022, doi: 10.1080/01605682.2020.1865846.
[13] M. E. Alqaysi, A. S. Albahri, and R. A. Hamid, "Evaluation and benchmarking of hybrid machine learning models for autism spectrum disorder diagnosis using a 2-tuple linguistic neutrosophic fuzzy sets-based decision-making model," *Neural Comput. Appl.*, vol. 36, no. 29, pp. 18161–18200, 2024, doi: 10.1007/s00521-024-09905-6.
[14] A. S. Albahri *et al.*, "Towards physician's experience: Development of machine learning model for the diagnosis of autism spectrum disorders based on complex T-spherical fuzzy-weighted zero-inconsistency method," *Comput. Intell.*, vol. 39, no. 2, pp. 225–257, Apr. 2023, doi: 10.1111/coin.12562.
[15] W. Nurhadi and S. Hasta, "Performance Comparison of Convolutional Neural Network with Traditional Machine Learning Methods in Adult Autism Detection," *Maisarah N. Appl. Technol. Comput. Sci. J.*, vol. VII, no. 1 SE-Articles, pp. 23–35, Dec. 2024.
[16] A. S. Albahri *et al.*, "Explainable Artificial Intelligence Multimodal of Autism Triage Levels Using Fuzzy Approach-Based Multi-criteria Decision-Making and LIME," *Int. J. Fuzzy Syst.*, vol. 26, no. 1, pp. 274–303, Nov. 2024, doi: 10.1007/s40815-023-01597-9.