

# Identifying Key Explainability Components in Black-Box AI Models for Emergency Department Disposition Prediction

Hananee Bueraheng<sup>1</sup>, Nurul Izrin Md Saleh<sup>1\*</sup>, Noraswaliza Abdullah<sup>1</sup>, Muhammad Iza Haris Fadhillah<sup>1</sup>, Maroning Useng<sup>2</sup>

<sup>1</sup>Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka (UTeM), Melaka, Malaysia

<sup>2</sup>Department of Business Administration (Business and Computer Science), International College, Thaksin University, Thailand

\*Corresponding Author

DOI: <https://doi.org/10.47772/IJRISS.2026.100900529>

Received: 24 September 2026; Accepted: 29 September 2026; Published: 15 October 2026

## ABSTRACT

Predicting patient disposition outcomes in the Emergency Department (ED)—specifically distinguishing between hospital admission and discharge home—is important for optimizing hospital capacity, staffing, and patient flow. While black-box Artificial Intelligence (AI) models such as Long Short-Term Memory (LSTM), K-Nearest Neighbors (KNN), Random Forest (RF), and Gradient Boosting Machines (GBM) achieve high predictive performance, their decision-making processes limit clinical interpretability and trust. This paper aims to evaluate and identify the key components of explainability across these four black-box models for ED disposition prediction. Using an analytical sample of 2,000 ED stays from MIMIC-IV-ED triage and demographic variables, we evaluated model performance across five standard classification metrics (Accuracy, Precision, Recall, F1-Score, and AUC) under an 80:20 stratified train-test split. Post-hoc feature attribution techniques—SHapley Additive exPlanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME)—were applied to identify dominant decision drivers and define core explainability dimensions. Random Forest achieved the highest predictive performance (Accuracy: 0.970, Precision: 0.963, Recall: 0.997, F1-Score: 0.979, AUC: 0.988), followed by KNN (Accuracy: 0.948, AUC: 0.977) and Gradient Boosting (Accuracy: 0.902, AUC: 0.944), whereas LSTM trained on single triage snapshots yielded lower performance (Accuracy: 0.875, AUC: 0.921). SHAP and LIME attributions highlighted clinician-assigned triage acuity (mean |SHAP| = 0.096), systolic blood pressure (0.065), body temperature (0.051), walk-in arrival mode (0.044), and oxygen saturation (0.042) as dominant decision factors. Synthesizing technical attributions with clinical requirements establishes five core explainability dimensions: Transparency, Interpretability, Fidelity, Consistency, and Actionability. Demonstrating predictive accuracy alone is insufficient for clinical adoption; establishing these grounded components provides a necessary foundation for human-centered decision support in emergency care.

**Keywords:** Explainable Artificial Intelligence (XAI); Emergency Department Disposition; Black-Box Models; SHAP; LIME; Feature Attribution; Clinical Decision Support Systems.

## INTRODUCTION

Emergency Department (ED) crowding is a major global healthcare challenge that adversely affects patient safety, treatment timelines, and operational efficiency. Patient disposition—the clinical decision determining whether an ED patient is admitted to an inpatient ward or discharged home—is a key operational node influencing hospital throughput and bed availability. Early and accurate prediction of disposition outcomes enables bed managers and clinical teams to proactively coordinate inpatient resources and streamline patient flow.

In recent years, Artificial Intelligence (AI) and Machine Learning (ML) algorithms have demonstrated strong capabilities in modeling non-linear interactions within clinical data. Black-box architectures such as Long Short-Term Memory (LSTM), Gradient Boosting Machines (GBM), K-Nearest Neighbors (KNN), and Random Forests (RF) can achieve high classification performance. However, their internal decision processes remain hidden from end-users. In high-stakes healthcare environments, this lack of transparency presents a substantial barrier to implementation, as clinicians and administrators require clear rationales to verify and trust model recommendations.

Post-hoc Explainable AI (XAI) techniques like SHapley Additive exPlanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME) assign feature attribution scores. Nevertheless, numerical attributions alone do not fully satisfy clinical reasoning requirements or explain how human reasoning can be systematically integrated into decision support. To bridge this gap, this study addresses the objective of determining the key components of explainability in AI black-box models (LSTM, Gradient Boosting, KNN, and Random Forest) for predicting Emergency Department (ED) disposition outcomes. Specifically, this paper evaluates the comparative performance of these four models on 2,000 ED stays, identifies dominant explanatory features using SHAP and LIME, and formalizes five multidimensional explainability components necessary for trustworthy clinical adoption.

## Related Work

### AI in Emergency Department Disposition & Patient Flow

Predicting hospital admissions and managing patient flow using machine learning has gained significant research traction. Studies demonstrate that routine triage data—including vital signs, arrival mode, demographic indicators, and acuity levels—provide strong predictive signals early in the patient journey. Ensemble tree-based methods like Random Forest and Gradient Boosting consistently demonstrate high performance on tabular clinical data due to their ability to capture complex feature interactions without strict parametric assumptions.

### The Black-Box Problem and Explainable AI (XAI)

Despite high accuracy, black-box models present severe transparency limitations in clinical workflows. Medical professionals must justify clinical decisions, making uninterpretable AI outputs difficult to adopt safely. Post-hoc attribution tools solve part of this issue: SHAP calculates unified feature attributions based on game-theoretic Shapley values, while LIME fits locally weighted linear surrogates around specific instance predictions. However, numerical feature attributions do not automatically form structured clinical rules or contrastive case explanations, highlighting the need for an explicit explainability framework.

### Dimensions of Explainability in Healthcare

Recent healthcare XAI literature emphasizes that explainability is a multidimensional construct comprising technical fidelity, cognitive interpretability, consistency, and actionability. In emergency care, explanations must align with clinician domain knowledge and fit time-sensitive workflows to prevent cognitive overload while ensuring patient safety.

## RESEARCH METHODOLOGY

The overall research framework consists of four sequential phases: Data Preparation, Black-Box Model Benchmarking, Post-Hoc Feature Attribution Analysis, and Synthesis of Core Explainability Components.

### Dataset Description and Pre-processing

The study utilized an analytical dataset of 2,000 ED stays derived from the MIMIC-IV-ED database structure. The dataset contains 1,429 admitted stays (71.5%) and 571 stays discharged home (28.5%). Ten primary predictors were analyzed: gender, arrival transport mode, body temperature, heart rate, respiratory rate, oxygen saturation (SpO<sub>2</sub>), systolic blood pressure (SBP), diastolic blood pressure (DBP), pain score (0–10 scale), and triage acuity level (ESI 1–5). One-hot encoding converted categorical variables into binary indicators (yielding

12 model predictors), and numeric triage predictors were standardized ( $z = (x - \mu) / \sigma$ ). The data were partitioned into an 80% training set (1,600 stays) and a 20% held-out test set (400 stays) using stratified sampling.

## Model Development & Performance Metrics

Four black-box architectures were trained under identical partitioning:

1. K-Nearest Neighbors (KNN): Instance-based similarity classification in standardized feature space.
2. Random Forest (RF): Bagged decision tree ensemble aggregating predictions across randomized feature subsets.
3. Gradient Boosting Machine (GBM): Sequential additive tree ensemble correcting residual errors.
4. Long Short-Term Memory (LSTM): Recurrent neural network evaluated on single-step triage vectors.

Model performance was measured on the held-out test set ( $n = 400$ ) using Accuracy, Precision, Recall, F1-Score, and Area Under the ROC Curve (AUC).

## Feature Attribution with SHAP and LIME

We derived global feature importance using mean absolute SHAP values across test instances. Local instance attributions were extracted using LIME by fitting locally weighted linear surrogates around target predictions.

## RESULTS AND DISCUSSION

### Comparative Predictive Performance

Table 1 presents the performance comparison of the four black-box models evaluated on the held-out test dataset.

Table 1: Performance Comparison of the Four Black-Box Models on Held-Out Test Data

| Model Family                  | Accuracy     | Precision    | Recall       | F1-Score     | AUC          |
|-------------------------------|--------------|--------------|--------------|--------------|--------------|
| K-Nearest Neighbors (KNN)     | 0.948        | 0.958        | 0.969        | 0.963        | 0.977        |
| Random Forest (RF) *          | <b>0.970</b> | <b>0.963</b> | <b>0.997</b> | <b>0.979</b> | <b>0.988</b> |
| Gradient Boosting (GBM)       | 0.902        | 0.897        | 0.976        | 0.935        | 0.944        |
| Long Short-Term Memory (LSTM) | 0.875        | 0.878        | 0.958        | 0.916        | 0.921        |

\* Selected reference black-box model for downstream explainability analysis.

Random Forest achieved the highest performance across all metrics, with an Accuracy of 0.970, Recall of 0.997, F1-Score of 0.979, and AUC of 0.988. Confusion matrix analysis (TP = 285, TN = 103, FP = 11, FN = 1) showed that Random Forest produced only a single false-negative admission prediction. KNN achieved the second-highest performance (Accuracy: 0.948, AUC: 0.977), confirming that local instance proximity in standardized vital-sign space provides strong predictive value. Gradient Boosting achieved 0.902 Accuracy and 0.944 AUC. LSTM recorded lower performance (Accuracy: 0.875, AUC: 0.921) because single triage snapshots provided no temporal sequences for the recurrent cell state to model over time.

### Key Feature Components Identified via SHAP and LIME

Feature attributions for the reference Random Forest model revealed the dominant decision factors governing ED disposition predictions.

Table 2: Key Feature Attributions Identified via SHAP for the Reference Model

| Predictor Variable            | Mean  SHAP | Direction & Clinical Impact                                                                      |
|-------------------------------|------------|--------------------------------------------------------------------------------------------------|
| Triage Acuity Level           | 0.096      | Dominant decision factor; high acuity (ESI 1–2) strongly drives admission.                       |
| Systolic Blood Pressure (SBP) | 0.065      | Primary vital sign; hypotension (<100 mmHg) increases admission risk.                            |
| Body Temperature              | 0.051      | Physiological indicator; fever ( $\geq 100.4^{\circ}\text{F}$ ) increases admission probability. |
| Arrival Transport Mode        | 0.044      | Operational proxy; ambulance arrival strongly correlates with admission.                         |
| Oxygen Saturation (SpO2)      | 0.042      | Respiratory metric; hypoxia (<95%) yields high admission likelihood.                             |

Secondary features (heart rate, respiratory rate, diastolic BP, pain score) provided smaller incremental attributions, whereas demographic variables like gender exhibited minimal influence.

### Multidimensional Explainability Framework

Synthesizing feature attributions with clinical requirements establishes five core dimensions of explainability for black-box ED disposition models:

1. **Transparency:** The ability to explicitly expose key input features, clinical thresholds, and operational factors influencing predictions.
2. **Interpretability:** The presentation of model reasoning in concise, human-readable formats that non-expert users can quickly grasp under emergency time constraints.
3. **Fidelity:** The mathematical agreement between the generated explanation layer and the underlying black-box model's true decision behavior.
4. **Consistency:** The stability of generated explanatory rationales across similar patient presentations, building clinical trust.
5. **Actionability:** The capacity of explanations to aid decision-making by showing how key clinical changes affect disposition outcomes.

### Accuracy vs. Explainability Trade-Off

Evaluating these four architectures highlights a clear trade-off between predictive performance and native model explainability. Complex ensembles like Random Forest achieve high accuracy but operate as opaque black boxes, whereas simpler models offer native interpretability at the expense of predictive power. Post-hoc XAI techniques resolve this tension by allowing high-performing black-box models to retain accuracy while providing auditable decision explanations.

## CONCLUSION

This study successfully accomplishes the objective by determining the key components of explainability across four black-box AI models (LSTM, Gradient Boosting, KNN, and Random Forest) for Emergency Department disposition prediction. Evaluated on 2,000 ED stays, Random Forest demonstrated superior predictive performance (Accuracy: 0.970, F1-Score: 0.979, AUC: 0.988). Feature attribution via SHAP and LIME isolated

triage acuity, blood pressure, temperature, arrival transport mode, and oxygen saturation as the primary decision drivers. Furthermore, this paper formalizes five essential explainability dimensions—Transparency, Interpretability, Fidelity, Consistency, and Actionability—to guide clinical decision-support development. These findings establish a conceptual foundation for subsequent symbolic knowledge representation in emergency care.

## ACKNOWLEDGEMENT

This work was supported by the Malaysia Ministry of Higher Education under the Fundamental Research Grant Scheme (Early Career Researcher) (FRGS/1/2024/ICT02/UTEM/02/14) and iDEAL Research Group of Universiti Teknikal Malaysia Melaka, UTeM. The authors also appreciate the valuable comments and suggestions provided by the anonymous reviewers and the editor, which helped improve the quality of this manuscript.

## REFERENCES

1. S. Pearce, T. Marchand, T. Shannon, H. Ganshorn, and E. Lang, "Emergency department crowding: An overview of reviews describing measures, causes, and harms," *Internal and Emergency Medicine*, vol. 18, pp. 1137–1158, 2023.
2. A. Brink, J. Alsma, L. van Attekum, W. Bramer, R. Zietse, H. Lingsma, and S. Schuit, "Predicting in-hospital admission at the emergency department: A systematic review," *Emergency Medicine Journal*, vol. 39, no. 3, pp. 191–198, 2022.
3. H. Elhaj, N. Achour, M. H. Tania, and K. Aciksari, "A comparative study of supervised machine learning approaches to predict patient triage outcomes in hospital emergency departments," *Array*, vol. 17, p. 100281, 2023.
4. Z. Liu, W. Shu, H. Liu, X. Zhang, and W. Chong, "Development and validation of interpretable machine learning models for triage patients admitted to the intensive care unit," *PLOS ONE*, vol. 20, no. 2, p. e0317819, 2025.
5. R. Roscher, B. Bohn, M. F. Duarte, and J. Garcke, "Explainable machine learning for scientific insights and discoveries," *IEEE Access*, vol. 8, pp. 42200–42216, 2020.
6. S. Choi, G. Kim, J. E. Kwon, and H. K. Lee, "Sequential anomaly detection for continuous prediction of unexpected ICU admission from emergency department," *IEEE Access*, vol. 12, pp. 96304–96318, 2024.
7. L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
8. J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *The Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.
9. J. J. Wadden, "Defining the undefinable: The black box problem in healthcare artificial intelligence," *Journal of Medical Ethics*, vol. 48, pp. 764–768, 2022.
10. C. Zednik, "Solving the black box problem: A normative framework for explainable artificial intelligence," *Philosophy & Technology*, vol. 34, no. 2, pp. 265–288, 2021.
11. S. Ali, T. Abuhmed, S. El-Sappagh, K. Muhammad, J. M. Alonso-Moral, et al., "Explainable artificial intelligence (XAI): What we know and what is left to attain trustworthy artificial intelligence," *Information Fusion*, vol. 99, p. 101805, 2023.
12. S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, vol. 30, pp. 4765–4774, 2017.
13. M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you?": Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, pp. 1135–1144, 2016.
14. T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial Intelligence*, vol. 267, pp. 1–38, 2019.
15. T. Tschoellitsch, P. Seidl, C. Böck, J. Meier, et al., "Using emergency department triage for machine learning-based admission and mortality prediction," *European Journal of Emergency Medicine*, vol. 30, no. 6, pp. 408–416, 2023.
16. A. Johnson, L. Bulgarelli, T. Pollard, L. A. Celi, R. Mark, and S. Horng, "MIMIC-IV-ED (Version 2.2)," *PhysioNet*, 2023.

17. L. Grinsztajn, E. Oyallon, and G. Varoquaux, "Why do tree-based models still outperform deep learning on typical tabular data?" in *Advances in Neural Information Processing Systems*, vol. 35, pp. 507–520, 2022.
18. A. J. Zeleke, P. Palumbo, P. Tubertini, R. Miglio, and L. Chiari, "Machine learning-based prediction of hospital prolonged length of stay admission at emergency department: A Gradient Boosting algorithm analysis," *Frontiers in Artificial Intelligence*, vol. 6, p. 1179226, 2023.
19. S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
20. R. J. Chen, J. J. Wang, D. F. K. Williamson, et al., "Algorithmic fairness in artificial intelligence for medicine and healthcare," *Nature Biomedical Engineering*, vol. 7, pp. 719–742, 2023.
21. F. Di Martino and F. Delmastro, "Explainable AI for clinical and remote health applications: A survey on tabular and time series data," *Artificial Intelligence Review*, vol. 56, pp. 5261–5315, 2023.
22. J. Jung, H. Lee, H. Jung, and H. Kim, "Essential properties and explanation effectiveness of explainable artificial intelligence in healthcare: A systematic review," *Heliyon*, vol. 9, no. 5, p. e16110, 2023.
23. C. I. Eke and L. Shuib, "The role of explainability and transparency in fostering trust in AI healthcare systems: A systematic literature review," *Neural Computing & Applications*, vol. 37, pp. 1999–2034, 2025.
24. Q. Xu, W. Xie, B. Liao, C. Hu, L. Qin, et al., "Interpretability of clinical decision support systems based on artificial intelligence from technological and medical perspective," *Journal of Healthcare Engineering*, vol. 2023, p. 9919269, 2023.
25. R. Rosenbacke, Å. Melhus, M. McKee, and D. Stuckler, "How explainable artificial intelligence can increase or decrease clinicians' trust in AI applications in health care," *JMIR AI*, vol. 3, p. e53207, 2024.
26. N. I. M. Saleh, H. Ab Ghani, and Z. Jilani, "Defining factors in hospital admissions during COVID-19 using LSTM-FCA explainable model," *Artificial Intelligence in Medicine*, vol. 132, p. 102394, 2022.
27. R. M. J. Byrne, "Counterfactuals in explainable artificial intelligence (XAI): Evidence from human reasoning," in *Proc. 28th Int. Joint Conf. Artificial Intelligence (IJCAI-19)*, pp. 6276–6282, 2019.