ML Models for Emergency Department Predictions

Dipesh Badal
TUD Dublin
X00229547@mytudblin.ie

Olohigbe Pearl Ohiwerei
TUD Dublin
X00229601@mytudblin.ie

John Staunton
TUD Dublin
X00229681@mytudblin.ie

Surya Teja Gowd Ayinavilli
TUD Dublin
X00229553@mytudblin.ie

ABSTRACT

The Emergency Department (ED) in a hospital is an extremely important, high-pressure and often resource-constrained environment. Patients are assessed and triaged for further care, with conditions varying from mild abrasions to life-threatening cardiac complications. With the recent advances in artificial intelligence and machine learning, there is an opportunity to apply these in the ED environment to gain new insights, make predictions and improve the quality of care.

This work develops and evaluates a set of machine learning models to predict whether an ED patient should be admitted to hospital for further care or sent home. The MIMIC-IV-ED dataset, which contains US patient data for approximately 425,000 ED visits, is used throughout. The models developed achieve AUROC scores in the 81% - 84% range which are comparable with those from other similar studies. The XGBoost and Deep Neural Networks models perform slightly better than Logistic Regression, Decision Tree and Random Forest models. However, lower Sensitivity scores are a source of concern and require further investigation.

This work also demonstrates how XAI techniques can be used to explain the predictions being made, and how Model Compression can reduce the model size and compute resources needed without having a major impact on accuracy. Ethical and legal compliance considerations are explored, along with an assessment of possible sources of bias and mitigation approaches.

CCS Concepts

• **Applied Computing** → Life and Medical Sciences → Health Informatics • **Computing Methodologies** → Machine Learning

Keywords

Healthcare, Artificial Intelligence, Bias, XAI, Model Compression, Feature Importance, SHAP.

1. INTRODUCTION

In Ireland, there were over 1.8 million attendances at hospital emergency departments in the past year (2024) [1], an increase of over 600,000 since 2015 [2]. Emergency Departments (ED) are a very important part of the healthcare system worldwide – they play a critical role in providing immediate care to seriously ill and injured patients, while also assessing which patients need to be admitted to the hospital for further investigations or treatments.

However, EDs are also high-pressure and resource-constrained environments. In Ireland, there are frequent media reports of long wait times [3] while earlier studies have highlighted ED factors that can impact on patient safety including the general working environment, levels of workload and the nurses' personal knowledge, experience and fatigue levels [4].

In this context, there is an opportunity for Artificial Intelligence (AI) to be applied as part of the patient assessment (also known as the "triage") process. Information that is available at the time of triage can be used to make a prediction of whether the patient will need hospital admission or can be discharged to home. This prediction can be made available for ED staff to take into account and could potentially help expedite necessary admissions while also avoiding mistaken discharges to the home.

A range of machine learning (ML) models can be applied to such a prediction task, which are often grouped depending on their level of explainability (glass-box, black-box and grey-box). Explanations can also be enhanced using XAI techniques such as LIME and SHAP to show which variables are most influential in a model's predictions. This use of AI in a medical setting also raises several ethical and legal compliance questions. To explore and address these topics in more detail, this paper has the following research goals and objectives:

1. Develop and compare glass-box and black-box deep learning models for predicting hospital admissions from the ED data.
2. Implement and evaluate explainable AI (XAI) techniques to enhance model transparency and understand which variables are having the most impact on predictions.
3. Explore emerging tools and techniques in healthcare AI, with a focus on 2 x separate model compression techniques.
4. Analyze ethical considerations and legal compliance issues specific to such emergency department AI applications.

2. LITERATURE REVIEW

2.1 AI in Hospital Emergency Departments

A systematic literature review by Boonstra & Laven in 2022 [5] found that the use of AI in Emergency Departments was still at a nascent stage. They reported that some studies suggested that AI-based tools have the ability to outperform human skills, but they concluded that the current technologies did not have the aims or power to do so. They added that AI-based tools can still provide valuable support with clinical decisions during triage and help relieve overcrowded ED's of their burden.

A more recent review by Tyler et al in 2024 [6] was more optimistic regarding the abilities of AI-based tools for ED – it found that ML models consistently demonstrated superior discrimination abilities compared to conventional triage systems, and that the integration of AI into the triage process yielded significant enhancement in predictive accuracy, disease identification and risk assessment. It also found that ML could accurately determine the necessity of hospitalization for patients in need of urgent care, and that ML improved resource allocation and quality of patient care.

2.2 MIMIC-IV ED Dataset - Background

The MIMIC-IV-ED dataset is part of the broader MIMIC-IV family of datasets which are intended to support a wide array of clinical research studies and related educational initiatives [7]. It is available to researchers on request via the PhysioNet team at the MIT Laboratory of Computational Physiology [8]. MIMIC stands for Medical Information Mart for Intensive Care and was first published back in 2006. MIMIC II followed in 2010, MIMIC-III in 2015 and the latest MIMIC-IV-ED used in this paper is Version 2.2 from January 2023.

2.3 MIMIC-IV ED Dataset - Uses

A recent study from Xie et al in October 2022 [9] has used the MIMIC-IV-ED dataset to undertake three clinical prediction tasks (1) prediction of hospitalization outcomes (2) prediction of critical outcomes (e.g. transfer to Intensive Care Unit or a mortality) (3) Prediction of ED re-attendance within 72 hours. For task (1) they achieved strong AUROC scores using a range of different ML models – Logistic Regression 80.6%, Random Forest 81.9%, Gradient Boost 81.9% and Multi-Layer Perceptron 82.2%. They found that more complex deep learning models (e.g. LSTM) did not perform better than the "simpler" models above. They also developed a re-usable data quality approach for handling min/max/outlier/missing values in Triage numeric data fields (e.g. for patient temperature or blood pressure) which used clinical expertise to set appropriate threshold values.

The majority of data in MIMIC-IV-ED is numeric or categorical in nature – but there is one free text field in the Triage table which records a patient's Chief Complaint. Xie et al developed an expert rule-based algorithm that looked for specific terms or phrases and then classified the complaint into one of 10 possible categories. McMaster et al [10] followed a similar categorization approach but then converted all variables into textual form, included descriptive column names and used a variety of BERT transformer models to make predictions which achieved AUROC score of 86.6%.

Lee et al [11] constructed a set of "clinical pseudo-notes" for each patient to also capture their ED visit details in text format. They used 5 BERT models pre-trained with clinical data to create embeddings for the Chief Complaint text, and used specialized foundation models for predictions, achieving AUROC scores >90% but also reported that their models did not generalize well when trained on MIMIC-IV and tested on Local data, and vice-versa.

2.4 Ethical, Legal & Regulatory Matters

The World Health Organization has developed a set of six key ethical principles for the appropriate use of AI for health [12] – (1) Protecting human autonomy (including privacy & confidentiality) (2) Promoting human well-being, safety & public interest (3) Ensuring transparency, explainability & intelligibility (4) Fostering responsibility & accountability (5) Ensuring inclusiveness & equity (6) Promoting AI that is responsive and sustainable. These align well with the seven requirements from the EU Assessment List for Trustworthy AI (ALTAI) [13], which in turn are based on the ethics guidelines from the EU High Level Expert Group (HLEG) on AI.

As most hospitals in Ireland are in the public sector, the Guidelines for the Responsible Use of AI in the Public Service published in May 2025 by the Irish government, are also highly relevant [14]. These are based on the same seven guiding principles from the EU HLEG and contain practical explanations and advice regarding each principle for public servants developing and using AI solutions. This report also provides guidance on how to adhere to the related legal requirements arising from key legislation including GDPR [15] and the EU AI Act [16], and highlights other relevant regulations to consider including the Data Governance Act, Digital Services Act and Digital Markets Act. It also includes an interesting mini case-study showing how an Irish government department has used AI to detect the presence of Personal Identifiable Information (PII) in grant applications, thereby helping the department to meet its own broader GDPR responsibilities.

2.5 Fairness / Bias in AI

While the ethical guidelines and regulations say that AI solutions need to be fair and avoid bias, they don't explain in detail how such biases can arise or how they can be detected and mitigated at an operational level. For the healthcare sector, Hasanzadeh [17] has undertaken a systematic review on bias recognition and mitigation strategies that apply in all stages of the AI model life cycle, from project conception right through to post deployment surveillance. He highlights that the dominant origins of bias in healthcare are human and can involve subconscious attitudes and stereotypes that have existed over long periods of time and are thus hard to detect and fix. Representation bias, in contrast, is easier to detect (e.g. an ML model may show varying levels of accuracy for individuals of a particular gender, ethnic group or age) and this type of bias can be mitigated through broader sample design, synthetic data generation and ML processing techniques [18]. It is also important to look out for intersectional biases where target groups overlap (e.g. female Asians), as highlighted by Byrne [19]

2.6 Explainable AI (XAI)

The ethical guidelines and regulations also say that AI solutions should be transparent and explainable – and recent research in the field of eXplainable AI (XAI) can help with this. Sajid Ali makes the distinction between "explainability" (why AI models make their decisions) and "interpretability" (how AI models make their decisions). He describes the challenges of XAI for black-box models (e.g. neural networks) in comparison with white-box models (e.g. logistic regression and decision trees). He also sees a category of grey-box models somewhere in between (e.g. random forest). It is also important to consider the need for XAI from different stakeholders' perspectives, as highlighted by Saeed's survey [20]. For example, the regulatory perspective may focus on the legal requirement for explanations while the end-user perspective may focus more on the need for trust and fairness.

Specifically in the healthcare sector, Subhan Ali's survey [21] investigates which XAI methods are being used most frequently – and finds that LIME and SHAP are the clear leaders appearing in 30% and 28% of publications respectively. Cynthia Rudin's advice, however, is to not keep trying to explain black-box models and instead to create models that are fully interpretable in the first place, especially when high-stakes decisions are being made [22]. She mentions decision tree and linear / logistic regression models as being highly interpretable. A further example is the Explainable Boosting Machine (EBM) model developed by Nori and Microsoft Research which is available on an open-source basis [23]. Showing how topics can be related, Cross et al [24] have highlighted the important role that XAI can also play in detecting and explaining Bias by examining how the features that drive predictions can differ for patients in different target groups.

2.5 2.7 Model Compression

As AI becomes more widely used in medical settings, there is growing interest in making such solutions as efficient as possible from a resource usage and computing power perspective. Model compression techniques such as Neural Network Quantization are gaining attention. Nagel [25] explains the two main classes of algorithms, namely Post-Training Quantization (PTQ) and Quantization Aware Training (QAT). In both cases, the weights and activation tensors are stored in lower precision than when the model was trained. This can result in some loss of model accuracy, but the gains come from reduced model sizes and faster inference times which can result in lower overall power consumption. PTQ is easier to implement as models do not need to be re-trained. Further implementation guidance is available from Google and TensorFlow covering all three options available (1) dynamic range quantization (2) float16 quantization (3) full integer quantization [26].

While model compression is widely considered for neural networks, some of the ideas and concepts can also be applied to other model types. For example. The Random Forest Classifier model has a max_depth parameter [27] which can be varied and can influence the accuracy, complexity and resulting sizes of models. This is, in effect, a form of model compression using a Pruning approach which may show benefits for this type of model.

3. EXPLORE & PRE-PROCESS DATA

3.1 Dataset Overview

The MIMIC-IV-ED v2.2 dataset includes six primary tables—edstays, diagnosis, medrecon, pyxis, triage, and vitalsign—linked via identifiers such as subject_id, hadm_id, and stay_id. Additional tables from the broader MIMIC-IV database—patients, admissions, and ICU stays—were incorporated using keys like patient_id and admission_id to enrich patient history and outcomes [10]. **Age Estimation:** To preserve privacy while approximating patient age at each visit, a derived feature, **stay_age**, was computed using anonymized fields from the *patient's* table. The formula used was: **stay_age = anchor_age + (year of ED admission − anchor_year)** This method, consistent with Xie et al. [10], provides a reliable estimate of patient age without revealing exact birth dates.

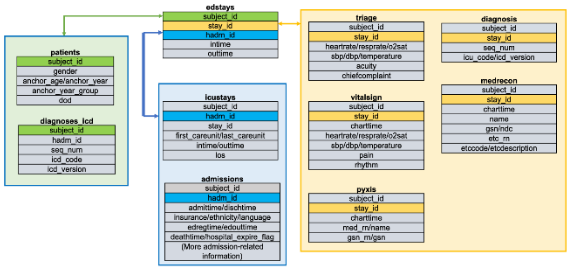


Fig 1: MIMIC-IV-ED Tables and Linking Keys [10]

3.2 Data Pre-processing Steps

The raw data underwent several cleaning procedures, including type conversions, null value handling, and outlier detection. Extreme values in clinical measurements were identified using expertly defined physiological thresholds and processed by following Xie's approach as follows:

- **< outlier_low or > outlier_high** → set to null
- **Between outlier_low and valid_low** → imputed to valid_low
- **Between valid_high and outlier_high** → imputed to valid_high
- **Within valid range** → retained as is

3.2.1 Feature Engineering

Several derived features were engineered for model performance:

- **Stay Time Hours**: ED admission minus discharge time
- **Stay Age**: calculated age at each ED visit
- **Previous ED Visits**: counts in the last 30/90/365 days
- **Medication Counts**: from **medrecon** and *pyxis* tables to reflect medication intensity

Chief Complaint Handling
Chief complaints recorded as free text were standardized into categorical classes to reduce sparsity and enhance model learning. Additional features, such as character and word counts, were derived for further analysis. Classification was performed using ChatGPT4o, as detailed in Appendix B. ChatGPT4o was chosen for several reasons including advanced NLP capabilities, contextual understanding, versatility, high performance and ease of integration.

3.2.2 Missing Data Management

Missing values in key variables (e.g., temperature, heart rate, respiratory rate, O_2 saturation, blood pressure, pain, acuity, chief complaint) were handled using statistical imputations (mean, median, or mode). Records with critical inconsistencies (e.g., disposition = ADMITTED but missing hadm_id) were excluded to ensure reliability.

3.2.3. Encoding and Transformation

Categorical variables such as gender, arrival_transport, race_group, and triage_chiefcomplaint_class were transformed via one-hot encoding for seamless integration into ML algorithms.
Additional Preprocessing
i. Temperature - converted from Fahrenheit to Celsius.
ii. disposition_ADMITTED - created as a binary target variable.
iii. Extensive data quality checks ensured the exclusion of rows with unresolved null values or inconsistencies.

3.2.4 Final Dataset Construction

A master dataset was created by merging the clean and processed tables. Preprocessing was conducted at the individual table level prior to integration to minimize error propagation and maintain data integrity. The final master data frame comprises 296,274 rows and 38 columns. This methodology aligns with the standard practices recommended in recent benchmark studies [10].

3.3 Bias Exploration

Demographic analyses were conducted to assess potential biases. Gender distribution appeared balanced, while most patients were White, followed by Black/African American and Hispanic/Latino. Age was concentrated in the 30–50 range. Common chief complaints included chest and abdominal pain. Vital signs showed near-normal distributions with some outliers. Slight variations in admission rates by gender and race were observed, informing future bias mitigation strategies during model training.

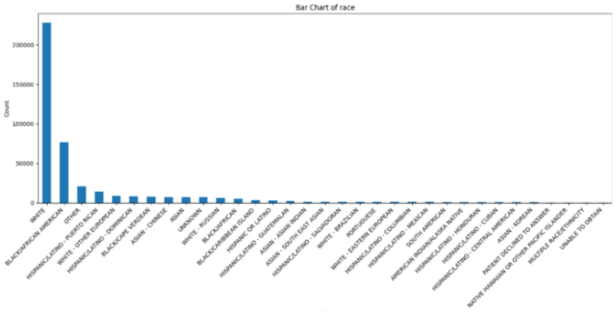


Fig 2: EDA on Race Column.

4. MODEL DEV & EVALUATION

To predict hospital admissions in emergency settings, we implemented and compared five supervised learning models using the cleaned MIMIC-IV-ED dataset. These included a Logistic Regression baseline and four advanced classifiers: DT, RF, XGBoost, and DNN. All tree-based models and the logistic regression were trained on a stratified 70–30 train-test split. In contrast, the DNN used a 70–15–15 split for training, validation, and testing to support hyperparameter tuning. Evaluation metrics included accuracy, precision, recall, specificity, and ROC AUC.

4.1 Logistic Regression

Logistic Regression served as a strong linear baseline with 75% accuracy and 0.82 ROC AUC, favoring specificity (84%) over recall (60%) for admitted patients, indicating a conservative approach to predicting admissions. Top features identified through correlation and SHAP included *triage_acuity*, *stay_age*, *medrecon_gsn_count*, and *arrival_transport_WALK IN*, with LIME and SHAP confirming the model's reliance on clinically intuitive inputs. Despite its simplicity, the model's transparency and interpretability make it a valuable benchmark.

4.2 Decision Tree / Random Forest

The Decision Tree model achieved its best performance with a max depth of 10, reaching 74% accuracy and 0.81 AUROC. Feature importance and SHAP analysis identified *triage_acuity*, *stay_age*, and *arrival_transport_WALK IN* as key predictors. However, visual interpretation became challenging with deeper trees due to float-based splits and complex branching.

Random Forest slightly improved performance to 75% accuracy and 0.83 AUROC using 100 trees with the same depth. It offered better generalization, with higher specificity (0.85) and precision (0.71), though recall remained moderate at 0.59. Feature importance remained consistent with DT, emphasizing clinically relevant variables like acuity, age, medication use, and past admissions.

4.3 XGBoost

XGBoost outperformed all tree-based models with 77% accuracy and an AUC of 0.84, effectively capturing non-linear feature interactions while maintaining strong generalization. SHAP identified *triage_acuity*, *stay_age*, and *medrecon_gsn_count* as top predictors, while LIME provided intuitive local explanations for both correct and incorrect classifications. Although the model showed strong overall performance, it demonstrated a clear imbalance between sensitivity (64%) and specificity (84%). This trade-off — favoring true negative predictions — was observed across all our models and warrants careful consideration, especially in clinical contexts where missing true positives may carry significant risk.

4.4 Neural Network

The deep learning model achieved 77% accuracy and 0.84 AUROC using a 3-layer feedforward network with ReLU activation and a sigmoid output, effectively utilizing key predictors after careful preprocessing, scaling, and handling of missing data. SHAP DeepExplainer revealed strong reliance on *triage_acuity*, *stay_age*, *arrival_transport*, and prior ED visits, with local explanations aligning well with clinical reasoning.

4.5 Model Evaluation

We compared models based on accuracy and per-class performance, including sensitivity (recall) and specificity for admissions (True class). Each model was further evaluated across gender, race, and age subgroups to identify any fairness gaps, although detailed fairness/XAI analysis is covered separately.
The table below summarizes standard classification metrics across all five models on the test set.

Table 1. Model Performance

Model	Acc.	Prec.	Recall / Sens.	Spec.	AUROC
LR	0.75	0.70	0.60	0.84	0.82
DT	0.74	0.68	0.59	0.83	0.81
RF	0.75	0.71	0.59	0.85	0.83
XGB	0.77	0.71	0.64	0.84	0.84
NN	0.77	0.72	0.62	0.86	0.84

XGB and NN achieved the highest accuracy at 77%, outperforming LR (75%), RF (75%), and DT (74%). However, all models showed lower recall for admitted cases (59–64%) compared to non-admitted ones, suggesting a tendency to under-detect high-risk patients. In an emergency department context, improving recall is crucial to avoid under-triage and ensure timely care for those needing hospitalization. From a performance perspective, XGB stands out due to its strong balance across metrics, robust generalization, and interpretability through SHAP and LIME. The NN is also a strong contender, especially in environments that can support the deployment of deep learning models.

5. XAI & FEATURE IMPORTANCE

To improve trust and transparency in clinical AI, we applied Explainable AI (XAI) techniques to interpret model predictions. **DT** and **LR** were treated as **glass-box models**, offering clear and direct interpretability. **XGBoost** and the **NN** were considered **black-box models** due to their complexity, while **RF** was classified as a **grey-box model**, providing some aggregated interpretability but lacking transparency at the individual decision level.

We used **SHAP** for global feature importance and **LIME** for local explanations of specific predictions, particularly to explore correct vs. incorrect classifications. For tree-based models, we also examined visualizations of decision paths to support interpretability analysis.

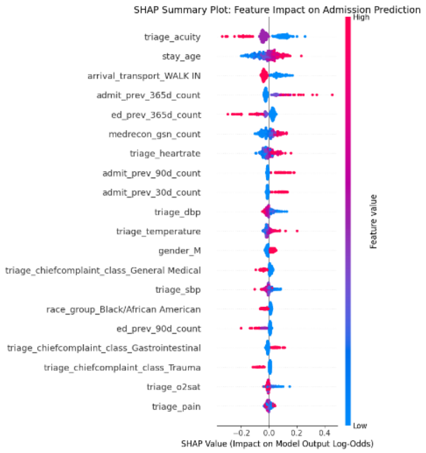


Fig 3: SHAP Plot from XGBoost Model

For illustrative examples of model predictions explained using LIME, including true positives, false positives, true negatives, and false negatives, please refer to Appendix D.

5.1 DT/RF

We used DT and RF for their interpretable structures. The DT with max depth 10 was hard to visualize due to complex float-based splits, so we reduced the depth to 5 for better readability with minimal performance loss. RF models showed similar patterns but lacked clear decision paths due to their ensemble nature. In both models, *triage_acuity*, *stay_age*, and *arrival_transport* were key features, aligning with clinical expectations. However, float-based thresholds (e.g., acuity ≤ 2.5) made it harder to translate decisions into clinical rules.

5.2 LR

We used SHAP for global interpretability and LIME for local explanations. SHAP showed that higher *stay_age*, lower *triage_acuity*, and increased medication use raised admission likelihood in the LR model. Locally, LIME explained both correct and incorrect predictions. True positives were influenced by high acuity and age, while false negatives often involved subtle arrival details or borderline vitals. The model's reasoning aligned well with clinical knowledge, making it the most transparent overall.

5.3 XGB

SHAP values for XGB showed very similar global patterns to LR and RF — top features included triage_acuity, stay_age, arrival_transport_WALK IN, and medrecon_gsn_count. SHAP summary plots revealed more nuanced, non-linear impacts. LIME analysis of correct/incorrect cases (for both admitted and home classes) confirmed good alignment: correctly admitted patients had low acuity and high med history, while false negatives often had borderline triage data. The model captured subtleties well but lacked full transparency, making post-hoc XAI essential.

5.4 NN

We used SHAP DeepExplainer for global analysis of the neural network, which highlighted top features like XGBoost—*triage_acuity*, *stay_age*, and *arrival_transport* were most influential. LIME confirmed that correct predictions were mainly driven by acuity and age, while incorrect ones often involved lower-impact features like *triage_chiefcomplaint*. Despite the NN's complexity, its alignment with XGB and LR boosted interpretability and trust.

5.5 Feature Importance

Combining top features from SHAP and model-specific feature importance functions showed strong consistency and stability.

Table 1. Model Performance

Feature	LR	RF	XGB	NN	Top 3?
triage_acuity	✓	✓	✓	✓	✓
stay_age	✓	✓	✓	✓	✓
arr_tran_Walk	✓	✓	✓	✓	✓
medrecon_gsn	✓	✓	✓	✓	X
admit_prev_90	✓	✓	✓	✓	X
triage_o2sat	✓	✓	✓	✓	X
ed_prev_365d	X	✓	✓	✓	X

Common Observations:

- All models highlighted acuity, age, and arrival mode — indicating strong clinical consensus.
- Medication history and previous usage appeared frequently but with slightly lower importance.

Differences:

- DT is harder to interpret globally due to tree complexity.
- NN explanations required SHAP & largely echoed XGB.

5.6 XAI / FI Summary & Discussion

All models consistently identified acuity, age, and medication burden as key admission predictors. Glass-box models like LR and DT offered better transparency, while black-box models (NN, XGB) required XAI to gain trust. From an explainability perspective, **LR** is most suitable for real-world deployment where clinical transparency is vital. However, **XGB** balances strong accuracy with reasonable interpretability using SHAP.

6. FAIRNESS & BIAS

Bias checks were undertaken by assessing accuracy of all models for target groups of gender, race and age group as outlined below.

6.1 Gender

Table 2 shows consistent performance of the models across both male and female patients. In all five models the female patients exhibit marginally higher accuracy by approximately 1% above the male patients. While the difference is very minimal, this consistent pattern may indicate subtle gender-based biases present in the data or during model training.

Table 2. Gender

Gender	DT	RF	LR	XGB	NN
Male	74%	75%	75%	76%	76%
Female	75%	76%	76%	77%	77%

Number of Instances: Female = 48746, Male = 40137

6.2 Race Group

Table 3 concerns disparities between racial groups. It reveals that the model performs higher for patients identified as Latino, Black, Asian or Unknown compared to their White or Other counterparts. For instance, NN achieves its highest accuracy for the Unknown group at 82%, while its lowest is for White and Other patients at 75%. These findings indicate underrepresentation or lower data quality for specific racial groups, particularly the patients categorized as White or Other.

Table 3. Race Group

Race	DT	RF	LR	XGB	NN
Asian	74%	79%	78%	80%	80%
Black	77%	78%	78%	79%	79%
Latino	78%	80%	79%	81%	80%
Other	74%	74%	74%	77%	75%
Unknown	79%	80%	81%	81%	82%
White	72%	73%	73%	74%	75%

N: AS=3889, BL=19944, LA=7253, OT=783, UK=4976, WH=52038

6.3 Age Group

Age-based disparities are clearly evident in the table below. Model performance is highest for younger patients (particularly those aged 18–30) and it progressively declines with increasing age. For example, the NN model achieves 87% accuracy for the 18–30 group but drops to 72% for the 75+ group. This trend is consistent across all models. This suggests a potential representation or complexity issue in older patient cases, which may require specific model refinement to improve results.

Table 4. Age Group

Age	DT	RF	LR	XGB	NN
18 – 30	86%	86%	86%	86%	87%
31 – 45	78%	78%	78%	79%	79%
46 – 60	71%	73%	73%	74%	74%
61 – 75	69%	71%	71%	72%	72%
75+	68%	70%	70%	71%	72%

N: 18-30=17045, 31-45=16127, 46-60=21623, 61-75=19108, 75+=14980

6.4 Intersectionality Bias

Intersectional bias checks were performed on the XGBoost model for target group gender with each of race group and age group. Some accuracy differences were observed (e.g. female 75+ age group accuracy was quite low at 70%)

6.5 Summary & Solutions to Overcome Bias

Overall, the results indicate a fairly balanced gender performance but reveal notable disparities across race and age groups. NN and XGB demonstrated the most consistent accuracy across subgroups, making them preferable when fairness is prioritized. Nevertheless, these models still underperform older individuals and certain racial categories, such as 'White' and 'Other'. The racial group 'Other' may potentially be influenced by its sample size. The reasons behind the lower accuracy for White patients remain unclear. Furthermore, continuous monitoring and fair evaluations are essential to ensure consistent, equitable performance across all patient populations, particularly in critical environments. Implementing strategies such as data balancing through oversampling, fairness-aware training, bias detection tools, Explainable AI (XAI), and the integration of fairness metrics (e.g., equal opportunity, demographic parity) can effectively mitigate biases and uphold ethical standards.

7. MODEL COMPRESSION

Model compression reduces the size and computational requirements of ML models through methods like quantization, pruning, and knowledge distillation. This is crucial for deploying sophisticated models on resource-constrained devices, enabling real-time processing, reduced latency, and lower energy consumption, which are essential for applications like autonomous vehicles, healthcare diagnostics and personalized assistants.

7.1 NN - Quantization

The implementation involved loading a pre-trained neural network model, preparing the data as NumPy arrays, and applying several quantization methods, including Dynamic Range, Float16, and Int8 Quantization. Each quantized model was saved and evaluated for accuracy, model size, and inference time. Dynamic Range and Float16 Quantization reduced model size and improved inference time while maintaining accuracy, whereas Int8 Quantization significantly reduced size & time but with a notable accuracy drop.

Table 5. NN Quantization Performance

Method	Acc.	Model Size (kb)	Infer Time (ms)
Baseline	76.6%	79.79	16.2422
Dynamic	79.2%	8.00	0.0079
Float16	79.2%	8.00	0.0063
Int8	34.8%	9.71	0.0053

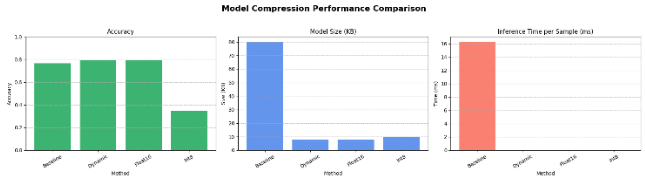


Figure 4. NN Model Compression Performance Comparison

Dynamic Range and Float16 Quantization improved accuracy marginally while reducing model size and inference times significantly. Int8 Quantization also reduced model size and inference time significantly but also dropped accuracy by 42% making it potential unsuitability for this type of use.

Quantized Model Usage

Quantized models are particularly useful in environments with limited computational resources, such as mobile devices, edge computing, and IoT applications. Their reduced size and faster inference times make them ideal for real-time applications where latency is critical. For instance, Dynamic and Float16 quantized models, which maintained high accuracy while significantly reducing size and inference time, can be deployed in mobile health monitoring apps to provide quick and efficient predictions.

7.2 Random Forest Model Compression

A form of model compression by pruning was implemented on the Random Forest model by varying the max depth parameter and assessing the impact on accuracy, model size and inference times as summarized below. Reducing the number of trees from 22 down to 10 has a negligible impact on accuracy but reduces model size by over 90% and inference time by over 50%.

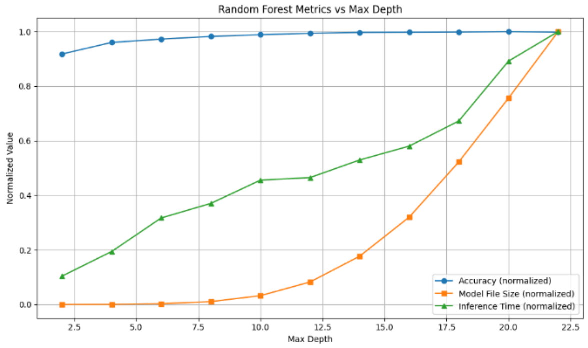


Figure 5. Random Forest Model Compression.

8. ETHICAL / LEGAL CONSIDERATIONS

8.1 Introduction

ED prediction applications using AI raise important ethical considerations and have related legal and regulatory requirements.

8.2 Ethical Considerations & Risks

To help identify and discuss the ethical considerations and risks, this paper is using the EU ALTAI requirements [13] as a checklist.

Human Agency & Oversight

An ED prediction and decision on hospitalization can quite literally be a matter of life and death. For this particular project, a "human-in-the-loop" approach should be followed, where the AI prediction is available to the ED staff as an input to their final human decision. There is a risk also of "automation bias" (where people place greater trust in decisions being made by technology ahead of human decisions) which clearly needs to be avoided in the ED setting.

Technical Robustness and Safety

This is important also, especially in the context of increasing levels of cyber-attacks against hospital IT environments. This application and all dev, test and production environments should be secured and managed in line with industry leading cyber security standards such as the NIST Cybersecurity Framework [28], SOC 2 Type II [29] and the NIS2 Directive [30] for Europe. Cyber-risks can also be mitigated through red-teaming and ethical hacking initiatives [31], leading to improved cyber resilience.

Regarding robustness and safety, we must consider ethically if a false positive (FP) is better or worse than false negative (FN). In this case, a FP means a patient is admitted unnecessarily with impact on hospital costs and patient inconvenience. However, a FN means a patient is sent home when they should be admitted to hospital for further treatment – which is a greater harm assuming we value health outcomes over costs and inconvenience.

Privacy & Data Governance

Privacy is a critical consideration for this type of project, as it deals with sensitive patient information. The MIMIC-IV-ED dataset owners have already removed any personal identifiable information (PII) from their tables, in line with their own HIPAA requirements. For similar projects in Europe, GDPR requirements on the capture, processing and retention of personal data must be followed. Data governance is important too in managing Privacy risks, and a Data Provenance Plan plays a key role in this.

Transparency

Transparency is a critical consideration for this type of project – full documentation on the datasets & models is needed, along with supporting XAI that can explain how and why the models are making their decisions. Regular and open communication with all impacted stakeholders is key – especially with regard to any limitations or assumptions in the AI system. Many AI projects are failing due to insufficient change management [32] for which transparency and communication are vital.

Diversity, Non-discrimination & Fairness

This is a critical consideration – as with all healthcare AI prediction solutions, they need to be accurate not just at the general population level, but also for specific target groups (e.g. gender, age, race). There are also risks of measurement bias (e.g. faulty heart rate monitors) and human bias (e.g. self-declared ethnicity). It is also important to monitor for bias as part of post-deployment surveillance.

Societal & Environmental Wellbeing

This is an important consideration – while this type of project can clearly help ED departments in their work, it is important to make such AI solutions available and accessible for all hospitals (not just the ones in rich countries with largest budgets). It is also important to consider energy and power consumption of such solutions, choosing models and compression techniques that can help in this regard. This in turn can support attainment of UN Sustainable Development Goals for "Industry, Innovation and Infrastructure" (Goal 9), "Responsible Consumption and Production" (Goal 12) and "Climate Action" (Goal 13).

Accountability

This is an important consideration – clear roles and responsibilities are needed right through the development lifecycle, additionally when the AI system goes live in production. The AI system should create audit logs and regular performance audits should be undertaken. A robust risk management system (e.g. the NIST AI RMF [33]) can help ensure this all happens.

8.3 Legal and Regulatory Compliance

Summary details regarding project compliance with legal / regulatory frameworks are included below.

EU AI Act

The AI application in this project maps to the "high risk" tier as it aligns very closely to the description for high-risk healthcare use cases in Annex 3 Section 5(a) of the Act. The organization developing the application will have "Provider" obligations that include creation of technical documentation, having a quality management system in place, undertaking conformity assessments and registering the application in an EU database. They will also need to collaborate with local market surveillance authorities. The Transparency obligations are also very important and may be the hardest ones to meet – while informing the user they are interacting with AI can be done via a user interface or report, the requirements for logging and audit trails of decisions are more challenging and latest XAI techniques do not (yet) provide a full solution here. The organization using the application will have "Deployer" obligations that include operating the system in accordance with instructions of use, ensuring human oversight of the system and ongoing monitoring and reporting of any serious malfunctions.

GDPR / HIPAA

GDPR applies for all patient data of citizens or residents of the EU and hospitals can find themselves playing both processor and controller roles. The key requirements include only capturing personal data needed for the specific prediction task, always keeping the data secure and private, undertaking Data Protection Impact Assessment (DPIA) and notifying any breaches to the relevant authorities and facilitating data subject requests to view or delete their data. For US patient data, the Health Insurance Portability and Accountability Act (HIPAA) [34] plays a similar role. The owners of the MIMIC-IV-ED dataset have already removed personally identifiable information (PII) from their tables as part of meeting their obligations here.

Others

For AI use in the public sector, the Irish government also recommends consideration of the Data Governance Act, the Digital Services Act and the Digital Markets Act [14].

8.4 Data Provenance Plan

A DPP is important for this project to describe the data sources & ownership, data quality, data legal compliance, data processing, data storage & security, data results reproducibility, data archiving & retention and post project data sharing – see Appendix A.

9. SUMMARY AND CONCLUSION

9.1 Results and Discussion

The predictive performances of the five models were very similar, with Accuracy in the 74% to 77% range and AUROC in the 81% - 84% range. The XGB and DNN models were slightly higher on both metrics than the LR, DT and RF models. This compares well with AUROC 81% - 82% from previous papers such as Xie [9]. While this may sound fine, we should recognize that 75% Accuracy is not actually a great score – it means that 1 in 4 outcome predictions are incorrect, so an ED nurse is unlikely to place a great deal of trust in such a system.

It is also worrying to see that our Sensitivity scores (~60%) are significantly lower than our Specificity scores (~mid 80%). The Xie paper had much more balanced results on similar models, reporting Sensitivity in the 75% - 76% range and Specificity in the 72% - 73% range. Our difference is due to our relatively high level of False Negatives, which means our predictions would send many patients' home who in fact should be admitted to hospital for further treatment. Lower accuracy is sometimes observed when the target class is significantly in the minority, but this is not the case here as our main dataset consisted of 38% admitted versus 62% sent home.

This lower Sensitivity could potentially be improved by (1) experiments on including further features in our models e.g. further triage data points or details of lab tests or other procedures performed while the patient was in the ED (2) further adjustments to model topologies and hyper-parameters (3) more sophisticated classification of the Chief Complaints text field (e.g. use a healthcare-specific LLM such as ClinicalBERT [35] to generate embeddings and include these as features in our models) (4) Develop a full LLM classifier model where the patient tabular data is serialized into a textual format and sent as a prompt to an LLM such as ChatGPT (similar to Lee's "pseudo-note" approach [11]).

Regarding overall Feature Importance, our models are all quite consistent and show that acuity, age, arrival transport, existing medications, triage vital signs and recent ED / hospital admissions all play a role. These are similar to the variables that Xie reported had the most impact on their Random Forest model.

The XAI analysis using SHAP (global level) and LIME (a local level) worked well, and their findings were broadly consistent with the Feature Importance analysis above. However, using tree images for the DT/RF/XGB models was challenging (1) GraphViz could not generate images of the full DT with 42 levels (2) Sklearn implementation of DT/RF does splitting of integers on decimal place boundaries which is confusing (3) XGB lowest level leaf in the tree image is an interim log probability which gets summed across all trees in the model – not very meaningful to an end user.

To assess bias, the performance of all models was calculated for the specific target groups of gender, race and age group. Overall performance was very similar for both genders, but somewhat lower levels of accuracy were observed for the white race group and the 61-75 and 75+ age groups. This can happen with ML models if certain target groups are under-represented – but that is not the case here as these groups make up 58%, 21% and 17% of the test data respectively. For completeness, checks for intersectional bias [19] were also performed for the combinations of gender / age group and gender / race group – and highlighted a somewhat lower accuracy for certain groups e.g. female 75+.

Regarding Bias, it is also important to recognize the role that Human Bias can play. For example, our Feature Importance and XAI analysis has shown that acuity is one of the most influential variables for all models. The acuity score is recorded by the ED staff, following the Emergency Severity Index (ESI) guidelines [36] – but human judgement and occasional error could arise. This type of bias cannot be detected or corrected by technical means alone, and mitigation requires broader staff training and quality management processes to be in place.

Model Compression showed that model size could be decreased, and inference time increased for both RF and DNN models without major impact on accuracy. This can reduce power and compute resources needed and will support UN Sustainable Development Goals 9, 12 and 13 [37].

Ethical considerations for this project were identified by using the requirements of the EU ALTAI as a checklist. All seven areas were found to be relevant and important for a project like this, and all are a source of potential risks if not closely managed. In addition to concerns over prediction accuracy and bias mentioned above, the further most important ethical concerns include (1) Human Agency & Oversight – it is critical that a "human-in-the-loop" approach be followed and that predictions are just an input to the final human decision on whether to admit a patient or not. (2) Privacy & Data Governance – also a critical ethical issue for a project like this. Patient data is highly sensitive and needs to be always managed in a secure and confidential manner, with the implementation of industry cyber standards (e.g. NIST CF) and a strong Data Provenance Plan playing important roles in this. (3) Transparency – is possibly the hardest ethical concern to resolve. While current XAI approaches can help, they do not fully explain the predictions that our AI solution makes, and they do not scale well if every single decision needs to have an audit trail (as the EU AI Act requirements for high-risk systems strongly suggests).

In addition to the EU AI Act, there are also legal compliance requirements regarding privacy and data subject rights arising from GDPR & HIPAA. The Data Governance Act, the Digital Services Act and the Digital Markets Act should also be considered. As ethical concerns and risks can arise in many areas and throughout all stages of an AI project lifecycle, the adoption of a comprehensive AI risk management framework (such as NIST AI RMF [33]) is also recommended.

9.2 Limitations

Limitations that have impacted on this paper include the availability of a single dataset (MIMIC-IV-ED). Being able to train and test models on another set of ED patient data could have produced a different set of results and findings.

9.3 Future Work

Future work could include efforts to improve the performance of the existing models as described above, including more sophisticated handling of the Chief Complaints text. It could also include the development of a full LLM classifier model. Given the challenges with XAI, future work could take on board the recommendations of Rudin [22] and seek to develop a fully interpretable model such as EBM using the InterpretML library in Python [23].

A final area for future work would be to get access to data for patients attending Irish hospital ED's and then train, test and compare model performance with this local data. Privacy requirements for such a study could be supported through the use of a Federated Learning approach [38].

10. REFERENCES

[1] "Hospital activity update - May 2025 - HSE.ie." Accessed: May 12, 2025. [Online]. Available: https://www.hse.ie/eng/about/who/acute-hospitals-division/hospital-activity/hospital-activity-update-may-2025.html

[2] "Key Trends: Hospital care." Accessed: May 12, 2025. [Online]. Available: https://assets.gov.ie/static/documents/chapter-3-hospital-care.html

[3] "Here's how long you can expect to wait in your local emergency department as new figures show scale of hospital delays | Irish Independent." Accessed: May 12, 2025. [Online]. Available: https://www.independent.ie/irish-news/health/heres-how-long-you-can-expect-to-wait-in-your-local-emergency-department-as-new-figures-show-scale-of-hospital-delays/a99684599.html

[4] Z. Fekonja, S. Kmetec, U. Fekonja, N. Mlinar Reljić, M. Pajnkihar, and M. Strnad, "Factors contributing to patient safety during triage process in the emergency department: A systematic review," *J Clin Nurs*, vol. 32, no. 17–18, pp. 5461–5477, Sep. 2023, doi: 10.1111/JOCN.16622;WGROUP:STRING:PUBLICATION.

[5] A. Boonstra and M. Laven, "Influence of artificial intelligence on the work design of emergency department clinicians a systematic literature review," *BMC Health Serv Res*, vol. 22, no. 1, pp. 1–10, Dec. 2022, doi: 10.1186/S12913-022-08070-7/FIGURES/2.

[6] S. Tyler *et al.*, "Use of Artificial Intelligence in Triage in Hospital Emergency Departments: A Scoping Review," 2024, doi: 10.7759/cureus.59906.

[7] A. E. W. Johnson *et al.*, "MIMIC-IV, a freely accessible electronic health record dataset," *Sci Data*, vol. 10, no. 1, pp. 1–9, Dec. 2023, doi: 10.1038/S41597-022-01899-X;SUBJMETA=174,228,308,478,692,700;KWRD=EPIDEMIOLOGY,HEALTH+SERVICES,PUBLIC+HEALTH.

[8] "About." Accessed: May 12, 2025. [Online]. Available: https://physionet.org/about/

[9] F. Xie *et al.*, "Benchmarking emergency department prediction models with machine learning and public electronic health records," *Sci Data*, vol. 9, no. 1, pp. 1–12, Dec. 2022, doi: 10.1038/S41597-022-01782-9;SUBJMETA=308,692,700;KWRD=HEALTH+CARE,MEDICAL+RESEARCH.

[10] C. McMaster, D. F. Liew, and D. E. Pires, "Adapting Pretrained Language Models for Solving Tabular Prediction Problems in the Electronic Health Record," Mar. 2023, Accessed: Mar. 27, 2025. [Online]. Available: https://arxiv.org/abs/2303.14920v1

[11] S. A. Lee *et al.*, "Emergency Department Decision Support using Clinical Pseudo-notes," Jan. 2024, Accessed: May 13, 2025. [Online]. Available: https://arxiv.org/pdf/2402.00160

[12] Organización Mundial de la salud, "Ethics and governance of artificial intelligence for health: WHO guidance," *OMS*, vol. 1, pp. 1–148, 2021, Accessed: May 13, 2025. [Online]. Available: http://apps.who.int/bookorders.

[13] "Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment", doi: 10.2759/791819.

[14] "Guidelines for the Responsible Use of AI in the Public Service." Accessed: May 13, 2025. [Online]. Available: https://www.gov.ie/en/department-of-public-expenditure-ndp-delivery-and-reform/publications/guidelines-for-the-responsible-use-of-ai-in-the-public-service/

[15] "EU General Data Protection Regulation (GDPR) - REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL," *official Journal of the European Union*, vol. 119, no. 1, 2016.

[16] "EU AI Act: EUR-Lex - 52021PC0206 - EN - EUR-Lex." Accessed: Dec. 14, 2024. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206

[17] F. Hasanzadeh, C. B. Josephson, G. Waters, D. Adedinsewo, Z. Azizi, and J. A. White, "Bias recognition and mitigation strategies in artificial intelligence healthcare applications," *npj Digital Medicine 2025 8:1*, vol. 8, no. 1, pp. 1–13, Mar. 2025, doi: 10.1038/s41746-025-01503-7.

[18] F. Hasanzadeh, C. B. Josephson, G. Waters, D. Adedinsewo, Z. Azizi, and J. A. White, "Bias recognition and mitigation strategies in artificial intelligence healthcare applications," *NPJ Digit Med*, vol. 8, no. 1, pp. 1–13, Dec. 2025, doi: 10.1038/S41746-025-01503-7;SUBJMETA=308,692,700;KWRD=HEALTH+CARE,MEDICAL+RESEARCH.

[19] A. Byrne, "The role of Intersectionality Within Algorithmic Fairness – HCAIM." Accessed: Feb. 14, 2025. [Online]. Available: https://humancentered-ai.eu/the-role-of-intersectionality-within-algorithmic-fairness/

[20] W. Saeed and C. Omlin, "Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities," *Knowl Based Syst*, vol. 263, p. 110273, Mar. 2023, doi: 10.1016/J.KNOSYS.2023.110273.

[21] S. Ali, F. Akhlaq, A. S. Imran, Z. Kastrati, S. M. Daudpota, and M. Moosa, "The enlightening role of explainable artificial intelligence in medical & healthcare domains: A systematic literature review," *Comput Biol Med*, vol. 166, p. 107555, Nov. 2023, doi: 10.1016/J.COMPBIOMED.2023.107555.

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

[23] H. Nori, S. Jenkins, P. Koch, and R. Caruana, "InterpretML: A Unified Framework for Machine Learning Interpretability," Sep. 2019, Accessed: May 13, 2025. [Online]. Available: https://arxiv.org/pdf/1909.09223

[24] J. L. Cross, M. A. Choma, and J. A. Onofrey, "Bias in medical AI: Implications for clinical decision-making," *PLOS Digital Health*, vol. 3, no. 11, p. e0000651, Nov. 2024, doi: 10.1371/JOURNAL.PDIG.0000651.

[25] M. Nagel *et al.*, "A White Paper on Neural Network Quantization," Jun. 2021, Accessed: May 13, 2025. [Online]. Available: https://arxiv.org/pdf/2106.08295

[26] "Post-training quantization | Google AI Edge | Google AI for Developers." Accessed: May 13, 2025. [Online]. Available: https://ai.google.dev/edge/litert/models/post_training_quantization#full_integer_quantization_of_weights_and_activations

[27] "RandomForestClassifier — scikit-learn 1.6.1 documentation." Accessed: May 13, 2025. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

[28] "The NIST Cybersecurity Framework (CSF) 2.0," Feb. 2024, doi: 10.6028/NIST.CSWP.29.

[29] "SOC 2® - SOC for Service Organizations: Trust Services Criteria | AICPA & CIMA." Accessed: May 16, 2025. [Online]. Available: https://www.aicpa-cima.com/topic/audit-assurance/audit-and-assurance-greater-than-soc-2

[30] "NIS2 Directive: new rules on cybersecurity of network and information systems | Shaping Europe's digital future." Accessed: May 16, 2025. [Online]. Available: https://digital-strategy.ec.europa.eu/en/policies/nis2-directive

[31] K. B. Chowdappa, S. S. Lakshmi, and P. N. V. S. Pavan Kumar, "Ethical Hacking Techniques with Penetration Testing", Accessed: May 16, 2025. [Online]. Available: www.ijcsit.com

[32] R. G. Cooper, "Why AI Projects Fail: Lessons From New Product Development," *IEEE Engineering Management Review*, 2024, doi: 10.1109/EMR.2024.3419268.

[33] "AI Risk Management Framework | NIST." Accessed: Mar. 28, 2025. [Online]. Available: https://www.nist.gov/itl/ai-risk-management-framework

[34] "HIPAA | HHS.gov." Accessed: May 16, 2025. [Online]. Available: https://www.hhs.gov/programs/hipaa/index.html

[35] K. Huang, J. Altosaar, and R. Ranganath, "ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission," Apr. 2019, Accessed: Apr. 02, 2025. [Online]. Available: https://arxiv.org/abs/1904.05342v3

[36] N. , T. T. , T. D. , & R. A. M. Gilboy, *Emergency Severity Index (ESI): A Triage Tool for Emergency Department Care, Version 4. Implementation Handbook 2012 Edition.* , 2012 Edition. AHRQ, 2012.

[37] "United Nationals Sustainable Development Goals (SDG)." Accessed: Dec. 15, 2024. [Online]. Available: https://sdgs.un.org/goals

[38] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," Apr. 10, 2017, *PMLR*. Accessed: May 18, 2025. [Online]. Available: https://proceedings.mlr.press/v54/mcmahan17a.html

[39] "What is the importance of a data provenance strategy? | Secoda." Accessed: May 14, 2025. [Online]. Available: https://www.secoda.co/blog/data-provenance-strategy

APPENDICES

Appendix A – Data Provenance Plan

Appendix B – ChatGPT Classification of Triage / Chief Complaints Text

Appendix C – Tree Images (for DT, RF, XGB models)

Appendix D – LIME XAI Images

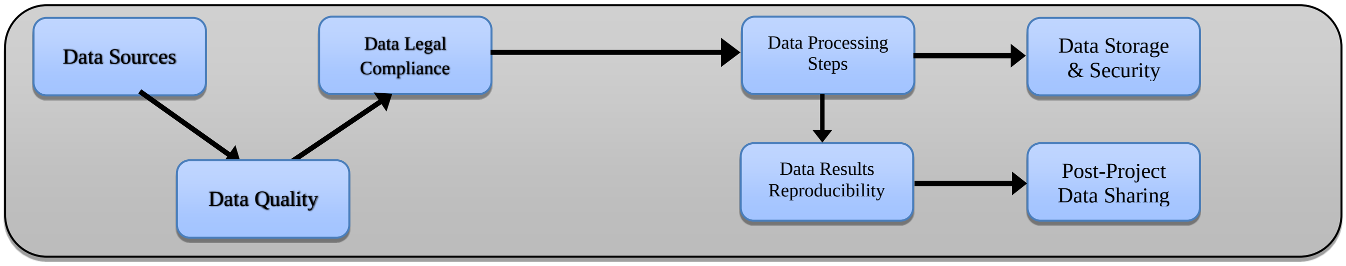
Appendix E – Individual Contribution Report

Appendix A – Data Provenance Plan

Data provenance involves comprehensive documentation of the origin, transformations, and overall lifecycle of data utilized within a machine learning pipeline. By maintaining a detailed provenance trail, organizations can ensure the reproducibility of their results, uphold accountability, and facilitate the detection of biases that may negatively influence model performance. Moreover, a robust data provenance framework is crucial for meeting regulatory compliance requirements, particularly in sensitive fields such as healthcare where data integrity and transparency are of utmost importance.

Our data provenance plan carefully highlights each phase of the data lifecycle, beginning with raw data ingestion and extending through various preprocessing steps, such as feature engineering, model training, and ultimately deployment. This method of rigorous documentation not only enhances transparency and trust in the machine learning process, but also establishes a strong basis for audits, compliance checks, and ongoing enhancements [39].

Stages	**Details**
01: Data Acquisition	- Source of data (MIMIC-IV-ED v2.2 from PhysioNet, BIDMC) - Date of extraction (15th of April 2025) - Access permissions/logs (Access to PhysioNet was approved and credentialed)
02: Raw Data Storage	- Storage location and format (Google Drive and OneDrive) - Integrity checks (e.g., hash values) - Backup history (CSV and IPYNB files are exported to OneDrive)
03: Data Cleaning & Preprocessing	- Cleaning scripts used (Handling missing values, standardizing codes) - Change logs (All were transformations logged) - Data quality issues
04: Feature Engineering	- Feature selection process - Transformation logic (Text vectorization & extraction of specific columns) - Normalization/scaling methods (Performed on certain columns)
05: Dataset Splitting	- Criteria for train/validation/test split (Stratified 70-30 split, admission vs discharge) - Random seed used (Set for reproducibility e.g., random_state = 42) - Stratification strategy (Performance split by gender, race, and age groups)
06: Model Training	- Data version used (Cleaned dataset v2.2) - Model version and parameters (Logistic Regression, DT & RF, XGBoost, NN) - Training logs (Automated logging of training metrics and weights with MLflow) - Fairness audit reports (Bias metrics recorded for subgroup analysis)
07: Model Evaluation	- Metrics recorded (Accuracy, AUROC, subgroup analysis, etc.) - Bias and subgroup analysis (Imbalances in demographics) - Explainability (SHAP and LIME results archived)
08: Deployment Data	- Dataset version used for inference (Versioned test split used for inference testing) - Drift monitoring setup (e.g., data drift, model decay) - Access logs
09: Data Storage & Security	- Full lineage records (Previous versions of datasets in OneDrive) - Archival timestamp (16th May 2025) - Data retention policy (Final versions of dataset and model were archived)
10: Data Legal Compliance	- GDPR, EU AI Act, and other relevant regulatory frameworks - Legal grounds for data processing (e.g., consent, public interest, research) - Data sharing agreements and mechanisms for data transfer
11: Data Results Reproducibility	- Version control of datasets and scripts with configuration management - Environment and dependency logging (Python 3.11, scikit-learn 1.4, TensorFlow 2.12) - Licensing, access control, and ethics review (16th May 2025)
12: Data Sharing	- Planned Sharing (Only certain results with no patient personal data - Exporting formats (PDF, IPYNB file, CSV files, etc.) - Licensing (Open-source MIT License for the codebase)



Appendix B – ChatGPT Classification of Triage / Chief Complaints Text

Model: ChatGPT 4o

Date: 29 April 2024

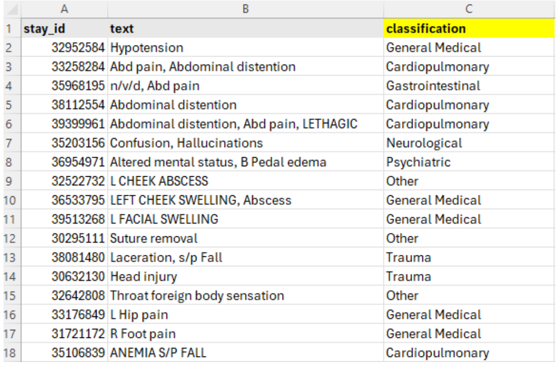
Input File: Triage.csv file from MIMIC-IV-ED. The two columns extracted and used were stay_id (numeric) and chief_complaint (text).

Prompt:

"I am a data science researcher and building a machine learning model to predict hospital outcomes for patents. Please use the Text column in the excel, which has a short summary of the patient's complaint, and classify it into one of 8 summary classifications. You should devise the classifications based on the content in the excel and your own knowledge of medical terms and conditions. Please add the new information as an extra column."

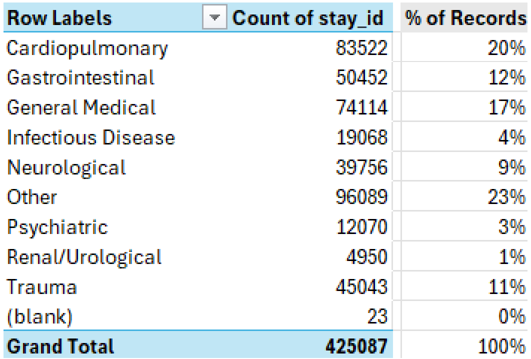
Output Results:

Csv file with extra classification column added in yellow, as shown below.



	A	B	C
1	stay_id	text	classification
2	32952584	Hypotension	General Medical
3	33258284	Abd pain, Abdominal distention	Cardiopulmonary
4	35968195	n/v/d, Abd pain	Gastrointestinal
5	38112554	Abdominal distention	Cardiopulmonary
6	39399961	Abdominal distention, Abd pain, LETHAGIC	Cardiopulmonary
7	35203156	Confusion, Hallucinations	Neurological
8	36954971	Altered mental status, B Pedal edema	Psychiatric
9	32522732	L CHEEK ABSCESS	Other
10	36533795	LEFT CHEEK SWELLING, Abscess	General Medical
11	39513268	L FACIAL SWELLING	General Medical
12	30295111	Suture removal	Other
13	38081480	Laceration, s/p Fall	Trauma
14	30632130	Head injury	Trauma
15	32642808	Throat foreign body sensation	Other
16	33176849	L Hip pain	General Medical
17	31721172	R Foot pain	General Medical
18	35106839	ANEMIA S/P FALL	Cardiopulmonary

Number of records per classification shown below.



Row Labels	Count of stay_id	% of Records
Cardiopulmonary	83522	20%
Gastrointestinal	50452	12%
General Medical	74114	17%
Infectious Disease	19068	4%
Neurological	39756	9%
Other	96089	23%
Psychiatric	12070	3%
Renal/Urological	4950	1%
Trauma	45043	11%
(blank)	23	0%
Grand Total	**425087**	100%

Appendix C – Tree Images (for DT, RF & XGB Models)

PDF files are attached separately.

Please note that wide tree images can be truncated when opened in Adobe Acrobat, so please open with a browser such as Microsoft Edge and use the zoom button (“+” and “-“) to see the full tree structure.

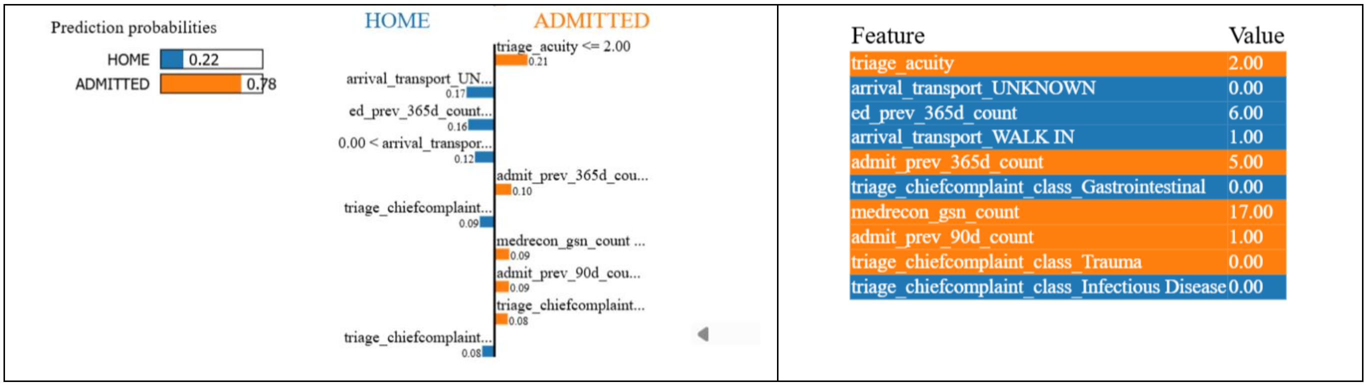
Model	**Filename**
Decision Tree, max_depth = 5	Decision_Tree_5.pdf
Decision Tree, max_depth = 10	Decision_Tree_10.pdf
Random Forest, max_depth = 10	Random_Forest_10.pdf
XGBoost	XGBoost_Tree_50.pdf Note – the 50 means that it is tree #50 out of 100 trees that the XGBoost model has created. It has no relevance for the number of levels in each tree.

Appendix D – LIME XAI Images

LIME XAI Examples (from the XGBoost model)

To illustrate how the XGBoost model made individual predictions, we examined four LIME explanations — one for each classification outcome: True Positive, False Negative, True Negative, and False Positive. Each case is summarized below.

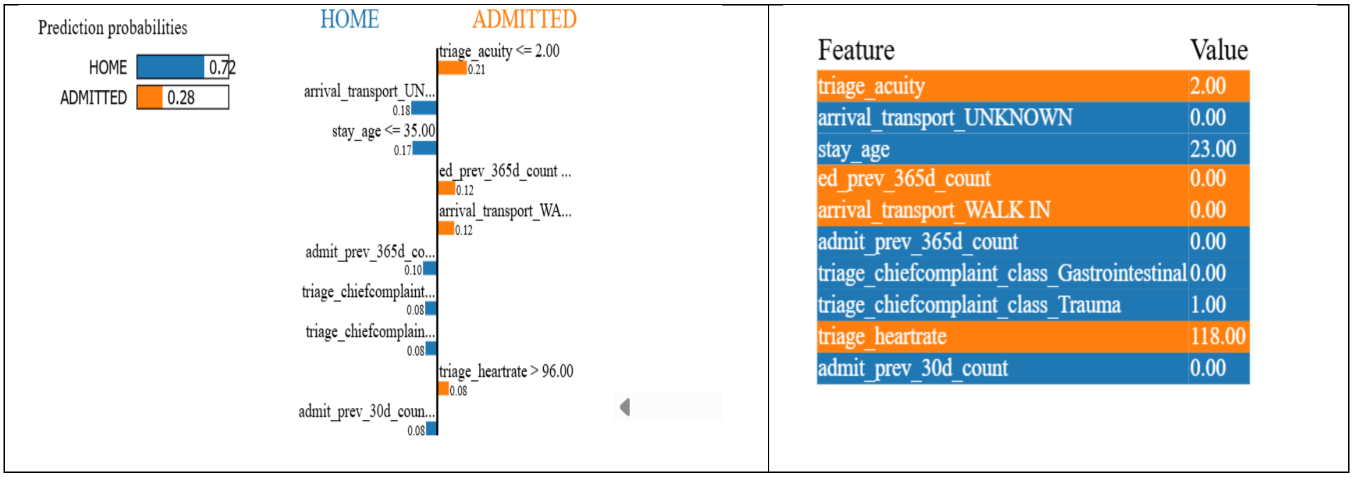
A.1 – True Positive (Correctly predicted ADMITTED)



Feature	Value
triage_acuity	2.00
arrival_transport_UNKNOWN	0.00
ed_prev_365d_count	6.00
arrival_transport_WALK IN	1.00
admit_prev_365d_count	5.00
triage_chiefcomplaint_class_Gastrointestinal	0.00
medrecon_gsn_count	17.00
admit_prev_90d_count	1.00
triage_chiefcomplaint_class_Trauma	0.00
triage_chiefcomplaint_class_Infectious Disease	0.00

In this case, the model predicted **ADMITTED** with a probability of 0.78, which matched the actual outcome. The patient presented with a **triage_acuity score of 2.00**, indicating moderate to high urgency. Additionally, the patient had a history of frequent healthcare use, including **six prior ED visits** and **five hospital admissions**, which signals ongoing or poorly controlled health conditions. The **medrecon_gsn_count was 17**, suggesting a substantial number of home medications and a high Medication disease burden. These combined features — clinical urgency, extensive medical history, and medication use — led the model to correctly classify the case as requiring admission.

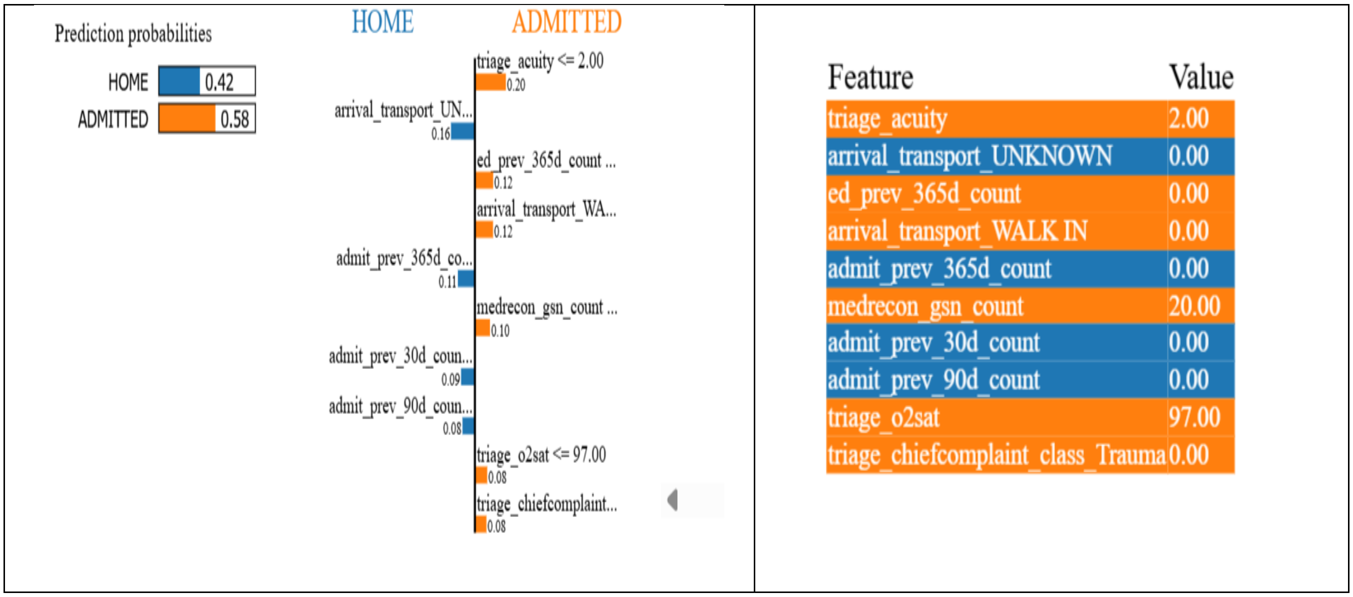
A.2 – False Negative (Incorrectly predicted HOME, actual = ADMITTED)



Feature	Value
triage_acuity	2.00
arrival_transport_UNKNOWN	0.00
stay_age	23.00
ed_prev_365d_count	0.00
arrival_transport_WALK IN	0.00
admit_prev_365d_count	0.00
triage_chiefcomplaint_class_Gastrointestinal	0.00
triage_chiefcomplaint_class_Trauma	1.00
triage_heartrate	118.00
admit_prev_30d_count	0.00

In this example, the model predicted **HOME** with a probability of 0.28, but the patient was in fact **ADMITTED**. The patient had a **triage_acuity score of 2.00**, suggesting moderate urgency, and a **heart rate of 118 bpm**, which is notably elevated and could indicate physiological distress. However, the patient was **only 23 years old** and had **no recorded history of prior ED visits or hospital admissions**. These missing historical indicators may have led the model to underestimate the admission risk, despite the presence of acute warning signs.

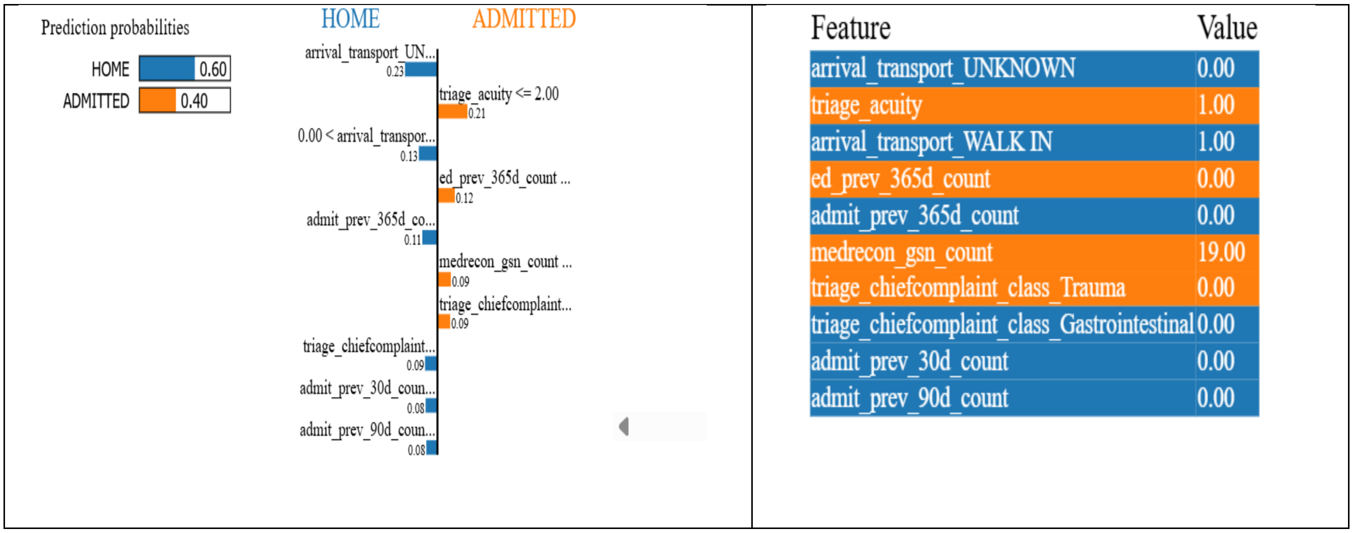
A.3 – True Negative (Correctly predicted HOME)



Feature	Value
triage_acuity	2.00
arrival_transport_UNKNOWN	0.00
ed_prev_365d_count	0.00
arrival_transport_WALK IN	0.00
admit_prev_365d_count	0.00
medrecon_gsn_count	20.00
admit_prev_30d_count	0.00
admit_prev_90d_count	0.00
triage_o2sat	97.00
triage_chiefcomplaint_class_Trauma	0.00

In this borderline case, the model predicted **HOME** with a probability of 0.42, which aligned with the actual outcome. The patient had a **triage_acuity score of 2.00**, indicating moderate urgency, and was taking **20 home medications**, suggesting the presence of one or more Medication health conditions. However, the patient had **no prior emergency department visits** and **no previous hospital admissions**, which implies limited recent interaction with acute care services. Despite the high medication burden, there were no other strong clinical signals pointing toward immediate risk.

A.4 – False Positive (Incorrectly predicted ADMITTED, actual = HOME)



Feature	Value
arrival_transport_UNKNOWN	0.00
triage_acuity	1.00
arrival_transport_WALK IN	1.00
ed_prev_365d_count	0.00
admit_prev_365d_count	0.00
medrecon_gsn_count	19.00
triage_chiefcomplaint_class_Trauma	0.00
triage_chiefcomplaint_class_Gastrointestinal	0.00
admit_prev_30d_count	0.00
admit_prev_90d_count	0.00

In this case, the model predicted **ADMITTED** with a probability of 0.60, but the patient was actually **discharged**. The key contributing features included a **triage_acuity score of 1.00**, which is the highest level of urgency on the acuity scale, and a **medication count of 19**, indicating significant health management. Despite these strong indicators, the patient had **no previous ED visits or hospital admissions**, suggesting this may have been an isolated acute episode rather than a continuation of Medication instability. The model likely overestimated the admission risk by heavily weighting the acuity and medication without sufficient context from past healthcare utilization.

Appendix E – STUDENT INDIVIDUAL CONTRIBUTION

This is available separately.