IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 8 May 2024, accepted 21 June 2024, date of publication 1 July 2024, date of current version 9 July 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3421295

RESEARCH ARTICLE

Predicting Hospital Stay Length Using Explainable Machine Learning

BELAL S. ALSINGLAWI[1], **FADY ALNAJJAR**[2], **MOHAMMED S. ALORJANI**[3], **OSAMA MOHAMMED AL-SHARI**[4], **MAURICIO NOVOA MUNOZ**[5], **AND OMAR MUBIN**[1]

[1]School of Computer, Data and Mathematical Sciences, Western Sydney University, Rydalmere, NSW 2116, Australia
[2]College of Information Technology, UAE University, Al-Ain, United Arab Emirates
[3]Department of Pathology and Microbiology, Faculty of Medicine, Jordan University of Science and Technology, Irbid 22110, Jordan
[4]Oncology Division, Department of Internal Medicine, Faculty of Medicine, Jordan University of Science and Technology, Irbid 22110, Jordan
[5]School of Built Environment, Western Sydney University, Rydalmere, NSW 2116, Australia

Corresponding author: Fady Alnajjar (fady.alnajjar@uaeu.ac.ae)

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the AAH Research Ethics Governance Committee, at Al-Ain Hospital under Approval No. AAHREC-09-20-027 via UAEU and the Western Sydney Research Ethics Committee under Approval No. H13511.

ABSTRACT Efficient bed management minimizes hospital costs and improves efficiency and patient outcomes. This study presents a predictive hospital-ICU length of stay (LOS) framework at admission, where it leverages hospital EHR. Our work utilizes supervised machine learning classification models to predict ICU patients' LOS in hospital clinical information systems (CIS). Our research marks the first known instance of utilizing explainable artificial intelligence (xAI) for the purpose of explainable machine learning applied to real data collected from hospital stays. We evaluated the predictive classification models using a range of performance metrics (Accuracy, AUC, Sensitivity, Specificity, F1- score, Precision, Recall and more) to predict short and long ICU lengths of stay upon hospital admission. XGBoost predicted short and long LOS with a 98 % AUC. This study shows how hospitals and ICUs might use machine learning to forecast patients on admission. Our study extends clinical information systems for hospitals to provide robust and trustworthy LOS, predictive models by using xAI to explain predictive model outputs.

INDEX TERMS Healthcare decision support systems, explainable artificial intelligence, machine learning, XGBOOST.

I. INTRODUCTION

The length of hospitalization serves as a common efficacy indicator in hospitals [1]. It significantly affects resource utilization and healthcare expenditures [2]. According to a report by the Australian National Health Performance Authority, shorter hospital stays are considered more efficient as they allow for the rapid availability of beds for new patients. However, unduly brief stays may compromise care quality and lead to adverse patient outcomes. Conversely, prolonged hospital stays, often resulting from complications, can heighten the risk of adverse health events. Delays in healthcare coordination, unrelated to the patient's clinical condition, may extend hospitalization durations. The report also noted that longer stays might result from delays in transitioning patients to other care services, such as aged care homes, community care services, or rehabilitation facilities [3].

Managing hospital bed availability and efficiency is crucial for addressing challenges in ICU, including patient overabundance, infections, mortality risk, and medical complications. To minimize these risks and improve resource utilization, a shorter ICU length of stay with high-quality care is necessary, especially in uncertain situations such as pandemics [4]. This not only lowers hospital charges but also ensures better outcomes for patients. Consequently, the availability of adequate bed spaces and timely patient transfers to other wards are critical for maintaining healthcare quality. Effective

The associate editor coordinating the review of this manuscript and approving it for publication was Asadullah Shaikh.

management of ICU resources is necessary to address these challenges and optimize healthcare delivery [5], [6], [7], [8].

ICU length-of-stay scores such as APACHE, SAPS, and SOFA [9], [10], [11], [12] are commonly used to estimate ICU resource utilization and predict mortality [13]. However, these approaches have limitations in estimating LOS and they are not disease-specific [14]. Therefore, research on accurate and trustworthy ICU resource consumption prediction systems is crucial. AI-based prediction systems using electronic health records (EHR) can provide more accurate and disease-specific approaches than existing systems. Linking database systems with clinical information systems (CISs) in the ICU potentially shorten ICU length of stays without affecting patient outcomes. Therefore, alternative ICU LOS assessment systems must be researched to optimize ICU resource utilization and improve patient outcomes [15], [16].

Clinical information systems (CISs) [15], including clinical decision support systems (CDSS) utilize electronic health record (EHR) data to improve healthcare delivery and control costs [16]. CDSS, in particular, is cost-effective and allows for effective use of EHR data [17], [18]. In recent years, research in non-knowledge-based CDSS systems has been initiated to utilize AI, machine learning (ML), and statistical learning to derive insights and patterns from data, making them a rapidly developing use case in hospital management and medicine. However, these systems have limitations, such as the "black-box" nature of the AI or ML used to make recommendations, and are not yet effectively incorporated into CDSS or CIS systems [19]. As the era of AI and data continues to transform clinical information systems, the future of personalized medicine (LDAPPM) appears to be promising for more effective and efficient healthcare delivery. Therefore, AI non-knowledge-based systems are not yet adopted and implemented in CDSS and generally within CIS systems [19], [20], [21], [22], [23], [24].

Our investigation builds on prior research that addressed ICU length of stay (LOS) as a binary prediction task. We review the most relevant studies in this area, including Ma et al. [25], who developed a tailored model using extreme learning machines (ELMs) and just-in-time learning methods (JITL) to predict ICU stays for personalized patient care. The combination of JITL and ELM achieved superior results compared to one-class SVM for predicting LOS of 10 days or more. However, the study did not examine the interpretability of the predictions.

Su et al. [26] compared XGBoost, Logistic Regression (LR), and Random Forest (RF) to SOFA for predicting ICU-sepsis patients' length of stay. They categorized LOS as (Short LOS: $\leq$ 6days, or Long LOS >6 days). They used the oversampling method SMOTE to treat imbalanced data. RF outdid XGBoost (AUC = 75%), LR (AUC 66%), and SOFA (AUC = 62%). The study was limited to the data gathering center, which led skewed predictions due to geographical characteristics. The study also focused on Sepsis and did not test the proposed models on other diseases in the dataset. Staziaki et al. [27] examined ANN and SVM models to predict LOS intensive care unit admission and extended LOS following torso trauma. They evaluated CT imaging (radiology reports), clinical characteristics such as (age, sex, vital signs, clinical scores, and laboratory values), and CT plus clinical qualities to predict LOS. Their findings reported that the combination of CT and clinical data properties (all features) improved prediction of both outcomes with ANN and SVM. The SVM model (all characteristics) predicted ICU-LOS admissions with (AUC = 87% ±0.03), whereas the ANN attained (AUC = 78% ±0.12). The study did not remove data noise and clean non-trauma patients, which can bias anticipated outcomes. The radiologist also adjudicates electronic radiological reports, therefore, interpretation biases may affect trauma patients' LOS expected outcomes.

Alghatani et al. [28] experimented with six classifiers to predict LOS (short: <2 days, long: >2 days). Using the MIMIC-III (v1.4) database, six classifiers (LR, RF, SVM, XGBoost, linear discriminant analysis: LDA, KNN: k-nearest neighbor) were tested on eligible ICU admissions [29]. A total of 33 features were used to predict the short and long LOS. Using quantiles, RF and XGboost outperformed other models (AUC = 69.78%, 69.69%). However, their system was limited to benchmarking the classifiers only on vital signs. Further, they did not explain the prediction decisions of the quantiles approach in an AI explainable approach. Gentimis et al. [30] used the MIMIC III database to predict length of stay (short LOS: < 5 days, long LOS: > 5 days) using ANN. They extracted 25 features from MIMIC-III tables that contained 25 features (admissions, CPT events, ICU stays, services, procedures ICD, and diagnoses ICD). ANN predicted LOS with 80% accuracy, however, the study lacks important model performance metrics such as (AUC, sensitivity, and specificity. These metrics are important to differentiate the model's performance in terms of accuracy and how likely the model is to distinguish the decision boundaries to effectively predict LOS short or LOS long. LOS prediction was performed using seven predictive models by Steele and Thompson [31]. In their work, the Bayesian Network (BN) achieved the best result among other predictive models with (AUC = 90%). However, the study suffered from drawbacks. For example, it did not specify the nature of clinical, laboratory, and vital signs collected to assess further models performance on more admission features considered a viable picture of the patient's information to identify the short from the long LOS.

Prior studies have examined different prediction models and settings to determine the most efficient model for predicting the length of stay in hospital environments. Ensemble learners are now widely used for predicting health outcomes and distinguishing between short and long LOS. Machine learning models, despite their predictive capabilities, often suffer from transparency, hindering their integration into clinical or administrative decision-making processes. There is a clear need for research that focuses on developing explainable AI techniques that are specifically designed for ICU systems in LOS-ICU contexts. Our study seeks to offer a framework for assessing hospital healthcare that is both predictive and explanatory. Our framework is specifically designed to be readily comprehended by AI non-experts. This versatility allows for more informed decision-making in both clinical and administrative settings. The objective of this framework is to enhance hospital workflow and resource utilisation by improving the transparency and interpretability of LOS predictions. In addition, it tackles a notable and previously unaddressed research issue in CIS, thus offering a valuable contribution to the progress of explainable AI in healthcare environments. Therefore, this study aims to contribute with the following:

- A proposed practical data-driven predictive framework for inpatient length of stay prediction in the ICU.
- A proposed model benchmarking technique to enhance LOS prediction and hospital resource utilization.
- A readily implementable framework for seamless integration into CIS prediction pipelines.
- A new, explainable prediction strategy for comprehensible outcomes for healthcare practitioners.

II. METHODS AND MATERIALS

A framework to predict the length of stay (LOS) for patients during their hospital admission, specifically their admittance into the ICU and discharge. This study use machine learning methods to predict the length of stay (LOS) of hospital inpatients using a real-world hospital dataset. Hence, this procedure is essential for assessing and validating prediction models using actual hospitalizations data. In this part, every step involved in the predictive framework (Fig 1) is addressed in detail. Therefore, the subsequent section describes each stage of the framework in detail.

A. DATA DESCRIPTION AND FEATURES EXTRACTION

Our retrospective study utilized electronic health record (EHR) data from Al-Ain hospital, encompassing all ICU admissions between December 31, 2017 and April 3, 2020 [58]. The de-identified nature of the EHR removed all patient details and identifiers in compliance with data protection regulations in the UAE and Australia. Our study population comprised 1045 distinct patients admitted to Al-Ain Hospital during the aforementioned period. Ethics approval was granted by the Al-Ain hospital and UAE University Ethics Committee (AAHEC-09-20-027), as well as preexisting amended ethics approval by Western Sydney University (WSU) with the ethics number (H13511).

This research employed a comprehensive inclusion protocol that covered all ICU hospitalizations at Al-Ain Hospital. Exclusion criteria involved expired hospitalizations and hospitalizations with significant missing data (Fig. 1). The International Classification of Diseases code ICD-10 [32] was used to classify diseases. The research framework included two experimental scenarios: a combined development set of all eligible patients (N = 1045) that included information from all four datasets, and a subset of patients' profile information in three separate sets. In total, 475 features were selected from the extracted electronic medical records at Al-Ain hospital. The inclusion criteria were jointly drafted by authors with medical and computer science backgrounds.

B. DATA PRE-PROCESSING AND DISCRETISATION

Data preprocessing is an essential task in the data mining process, particularly in Electronic Health Record (EHR) datasets, which often suffer from missing values, outliers, or raw data that require further processing and feature redundancy [33]. This study aimed to process and extract features that contributed to the patient's stay at Al-Ain Hospital. Several steps were performed during the preprocessing stage, which is similar to those carried out in previous works [34], [35].

One of the primary challenges in handling datasets with many missing values, non-values (NaN), or blanks is training machine learning models that can drastically impact the machine learning quality, performance, and predicted outcomes [36]. To address this challenge, data imputation was used to handle the missing values or values containing blanks in the four imported tables of the Al-Ain dataset [36]. The null function from the Pandas library in Python was used to replace any non-value with a zero value (0) since the input was not available or possible due to the non-applicable option. Furthermore, categorical variable transformation is necessary, especially when dealing with nominal or categorical variables, such as the Al-Ain dataset with several nominal attributes that require further data representation [36]. Therefore, the one-hot encoding method was used to transform categorical attributes into nominal and binary attributes, which improves the performance of machine learning models [37].

Data discretization involves transferring numeric or continuous variables into nominal or categorical variables with minimal loss of information. Statistical studies have examined the rationale behind data discretization and proposed methods to transfer continuous variables into nominal or categorical variables [38], [39]. For example, in electronic health records (EHR) and in clinical decision support systems (CDSS) studies [38], [40] they binned the continuous variables into nominal target variables. In hospital CIS systems, binning the (continuous) length of stay into nominal and categorical is accompanied by advantages for healthcare caregivers to maximize hospital resource utilization [41], [42]. In this study, the continuous variable length of stay (LOS) was binned into a binary class LOS approach based on previous studies, which grouped short LOS to (0-7 days) and long LOS to (>7 days) [43], [44], [45]. Therefore, the LOS was discretized into two labels: label zero (0) for a short length of stay (0-7 days) and label one (1) for a long length of stay (7+ days), resulting in a predictive LOS task with a classification problem.

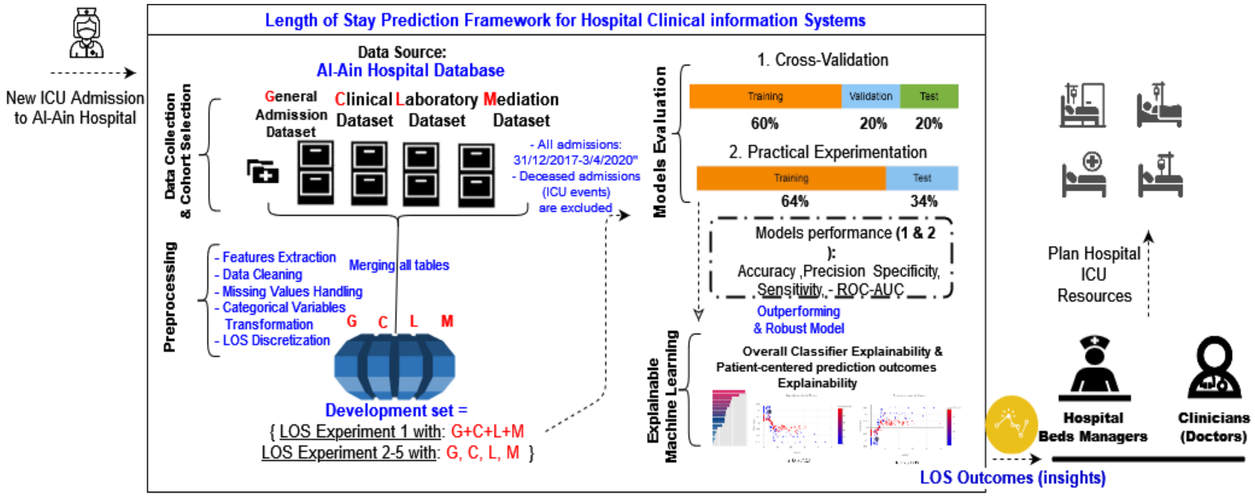

Length of Stay Prediction Framework for Hospital Clinical information Systems
New ICU Admission to Al-Ain Hospital
Data Source:
Al-Ain Hospital Database
Data Collection & Cohort Selection
General Admission Dataset
Clinical Dataset
Laboratory Dataset
Mediation Dataset
- All admissions: 31/12/2017-3/4/2020"
- Deceased admissions (ICU events) are excluded
Preprocessing
- Features Extraction
- Data Cleaning
- Missing Values Handling
- Categorical Variables Transformation
- LOS Discretization
Merging all tables
G C L M
Development set =
{ LOS Experiment 1 with: G+C+L+M
LOS Experiment 2-5 with: G, C, L, M }
Models Evaluation
1. Cross-Validation
Training
Validation
Test
60%
20%
20%
2. Practical Experimentation
Training
Test
64%
34%
Models performance (1 & 2):
Accuracy ,Precision Specificity, Sensitivity, - ROC-AUC
Outperforming & Robust Model
Explainable Machine Learning
Overall Classifier Explainability & Patient-centered prediction outcomes Explainability
Plan Hospital ICU Resources
Hospital Beds Managers
Clinicians (Doctors)
LOS Outcomes (insights)

FIGURE 1. Predictive LOS framework on ICU hospitalisations from real hospital dataset.

C. MODELS SELECTION AND PERFORMANCE EVALUATION

This section evaluates the machine learning models [58] used to assess the predictive LOS framework on ICU hospitalizations using real hospital datasets. The implementation, tuning, and performance evaluation models used Python and the Sklearn.

1) MACHINE LEARNING ALGORITHMS

a: THE EXTREME GRADIENT BOOSTING (XGBOOST)

The eXtreme Gradient Boosting (XGBoost) algorithm [46] is an ensemble-based learning (boosting) model. The XGBoost implements gradient boosted decision trees [47] designed for performance and speed. A recent implementation of the gradient tree boosting machines involves combining the predictions of many "weak learners" of decision trees into a strong predictor. In addition, it uses more regularised model formalization to control the overfitting and give it better performance [46]. One of XGBoost's significant advantages is that it is designed for scalable datasets. We used learning rates (0.01, 0.1, 1), the number of estimators (5, 50, 250), and the maximum depth of (1, 3, 4, 5,9) for the hyperparameters in the cross-validation stage. Table 1 describes the hyperparameters' values (All Dataset features, G, C, L, and M).

b: RANDOM FOREST

random forests (RF) is an ensemble learning method (bagging) for classification that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification). Each time we build a model based on the decision tree trained on row/features sampling with replacement. Every time we build the decision tree model, we have new rows fed into the new decision tree learners (bootstrapping). The bootstrapping process occurs in parallel until we achieve (n) of models trained on the decision tree. Eventually, we aggregate models that are generated from bootstrapping using majority voting to give the final predictive output. In the context of our study, the voted majority of the RF is (0: Short LOS or 1: Long LOS). We input the RF classifier with a decision tree as the base learner, consisting of up to 250 trees with a Gini index and a maximum depth of 21. Table 1 donates all hyperparameters values (All Dataset features, G, C, L, and M) using the RF classifier.

c: GRADIENT BOOSTING MACHINES

Gradient Boosting (GB) is a powerful ensemble learning technique (boosting) for building predictive models [47]. It works by producing a prediction model from an ensemble of weak prediction models like decision trees. It creates new base learners to be maximally correlated with the negative gradient of the loss function and associated with the whole ensemble. Therefore, it builds the model (a weak learner), and it improves model errors over time. It achieves its best performance over a sequential process after training and learning, and eventually, we get an improved model with better predictive outcomes. We used learning rates of (0.01, 0.0, 1, 10,100) with a number of trees of (5, 50, 250, and 500) as well as the max depth of (1, 3, 5, 7, 9). We used GirdSearch (cross 5-fold validation) and attained the GB's hyperparameters values and setups according to experimental sets (All Dataset features, G, C, L, and M) as described in Table 1.

d: LOGISTIC REGRESSION (LR)

logistic regression is a statistical method based on the use of a logistic function (sigmoid function) to model the output of binary values (0 or 1) [47]. The logistic regression model used (L1, L2, and elasticnet) as the regularisation (penalty) or no regularisation input. In addition, we used solvers

TABLE 1. Hyperparameters of the predictive models.

Parameters	Description	Experiment setup				
		All - G+C+L+M	G	C	L	M
	Logistic Regression (LR)					
C = 1	To control penalty strength (Inverse of regularization strength), and it must be a positive value.	1000	C=1	C=0.001	C=0.01	C=1
Solver = [liblinear, newton-cg]	for regularization (penalty) and optimization problem.	N/A	liblinear	newton-cg	liblinear	liblinear
	Multi-layer Perceptron (MLP)					
hidden_layer_sizes	Describes the ith element represents the number of neurons in the ith hidden layer.	10	50	50	N/A	10
activation	Refer to activation function for the hidden layer.	logistic	N/A	logistic	logistic	tanh
learning_rate	Learning rate schedule for weight updates.	0.01	0.01	0.01	0.01	0.01
	Random Forest (RF)					
n_estimators	Describes the number of trees in the forest.	50	250	5	250	50
max_depth	Describes the maximum depth of the tree.	None	8	4	8	8
max_features	Describes the number of features to consider when looking for the best split.	None	sqrt	log2	log2	sqrt
	Gradient Boosting (GB)					
n_estimators	Describes the number of boosting stages to perform.	500	50	5	50	500
max_depth	Refers to the maximum depth limits the number of nodes in the tree.	1	9	1	5	N/A
learning_rate	Learning rate shrinks the contribution of each tree	0.01	0.01	0.01	0.01	0.01
	eXtreme Gradient Boosting (XGBoost)					
n_estimators	Describes the number of gradients boosted trees (equivalent to the number of boosting rounds)	100	5	5	5	250
max_depth	Describes the maximum tree depth for base learners	None	3	1	5	3
learning_rate	Describes the Boosting learning rate	None	1	0.01	0.1	C=1

(newton-cg, lbfgs, liblinear), and the inverse of regularisation strength (C) as a positive float value. The GirdSearch with cross 5-fold validation obtained the LR's hyperparameter values and setups per the experimental sets (All Dataset features, G, C, L, and M) as described in Table 1.

Algorithm 1 Prediction Algorithm for Baselining Stage (1): Cross-Validation With Hyperparameters

```
Require:
    $K$: number of k-fold, $K = [k_1, k_2, \ldots, k_5]$;
    Target label: Short, OS, Long OS;
    $D$: transformed dataset (e.g., EHR), containing input features $X$, and output feature $y$;
    $H$: set of hyperparameters $H\_sets$ with different values;
    $M$: set of models $M = [\text{XGBoost, RF, GB, LR, MLP}]$;
Ensure: Set of outperformed models matrix $M_P$, performance estimation on $V_{j,validate}$ vs. $E_{j,test}$
procedure CrossValidationWithHyperparameters($K, D, H, M$)
    for $i = 1$ to $K$ do
        Split $D$ into $D_{train}, D_{test}, D_{validate}$ for the $i$-th split
        for $j = 1$ to $D_{train}, D_{test}, D_{validate}$ do
            for each $h$ in $H_{sets}$ do
                Train $M$ on $D_{train}$ with hyperparameter set $h$
                Compute test error $E_{j,test}$ for $M$ with $D_{test}$
            end for
            Select optimal hyperparameter set $h$ from $H_{sets}$
            Train $M$ with $D_{train}$ using $h$
            Compute test error $E_{j,test}$ for $M$ with $D_{test}$
            Compute validation error $V_{j,validate}$ for $M$ with $D_{validate}$
        end for
    end for
    return Set of outperformed models matrix $M_P$, performance estimation on $V_{j,validate}$ vs. $E_{j,test}$
end procedure=0
```

e: MULTI-LAYER PERCEPTRON NEURAL NETWORK (MLP)

The MLP [48] is a machine learning predictive model that mimics the neural networks stimulated by the biological neural networks and solves challenging computational tasks such as predictive modelling tasks. We used a feedforward, multi-layer perceptron neural network comprising three hidden layers with 10, 50, and 100 neurons. The activation functions are Relu, Tanh, and Logistic. We trained the network on three learning rates (constant, invacaling, and adaptive). All MLP's hyperparameters values for the five experimental sets (All Dataset features, G, C, L, and M) are described in Table 1.

2) THE HYPERPARAMETERS OF THE PREDICTIVE MODELS

Hyperparameter selection was implemented as model-based. A Grid-Search strategy with cross 5-fold cross-validation was used to find the hyperparameters used to get good predictive results in the binary approach. This step is essential in practise and experimental settings to allow tailing the behaviour of machine learning models, especially in the context of this study (the electronic medical records dataset "Al-Ain hospital").Table 1 discusses each model with its hyperparameter values and explanation.

Algorithm 1 details a cross-validation procedure for baselining with hyperparameters, involving multiple models and hyperparameter sets. The algorithm iteratively splits the dataset, trains models with different hyperparameter configurations, and evaluates their performance. It identifies the optimal hyperparameters for each model based on validation errors, ensuring the selection of the best-performing models for further analysis. Whereas, Algorithms 2 presents the steps involved in the evaluation of the outperforming models from stage (2).

3) MODELS EVALUATION METRICS

A set of evaluation metrics was utilized to evaluate classifiers in the predictive LOS framework of ICU hospitalizations from real hospital datasets. In the first stage (models' benchmarking stage), cross-validation with k-fold = 5 was implemented to estimate the skills of proposed classifiers on unseen data. The metrics used in this s metrics, such as Accuracy, Precision Sensitivity, Specificity and AUC, are used to assess the classifiers' performance in predicting the short and long LOS for actual admissions Al-Ain hospital dataset. Also, we used the statistical measure (confidence interval: CI). Therefore, the performance evaluation metrics are donated as follows:

a: ACCURACY

donates the ratio of the correct predictions to the total of a number of predictions:

$$Accuracy = (TP + TN)/(TP + TN + FP + FN) \qquad (1)$$

b: PRECISION

refers to the number of positive classifications that are actually correct (or called "positive predicted value" or "PPR"):

$$Precision \ = TP/(TP + FP) \qquad (2)$$

c: SENSITIVITY (RECALL)

measures the proportion of actual positives that are well classified (or called the true positive rate, or TPR, or Recall):

$$Sensitivity = TP/(TP + FN) \qquad (3)$$

d: SPECIFICITY

measures the proportion of actual negatives that is well classified (true negative rate 'TNR').

e: F1-SCORE

It can be interpreted as the weighted average of precision and recall. F1-Score = 1 is the best possible value, and F1-Score close to 0 is the worst value.

$$F1 - \text{ Score } = 2 * \frac{\text{Preision } * \text{ Recall}}{\text{Precision } + \text{ Recall}} \qquad (4)$$

f: AREA UNDER THE ROC CURVE (AUC)

The AUC measures the quality of the model's predictions regardless of what classification threshold is chosen. It represents the area under the ROC curve plots (TPR vs. FPR), where TPR is the true positive rate, and FPR is the false positive rate. The ROC plot visualizes the tradeoff between the classifier's sensitivity and specificity.

g: PR-AUC

The (area under the precision-recall curve) PR-AUC is a curve that combines precision and recall in one plot (single visualization). Thus, once we calculate precision and recall for every threshold, the higher the y-axis curve is, the better the model performance. Therefore, the optimal operating point on PR curve n a PR curve is the upper right corner, and the values of PR-AUC range from 0 to 1, with a note that 1 describes a perfect classifier [49].

h: K-FOLD "CROSS-VALIDATION"

Cross-validation is a statistical method that is used to estimate the skill of a machine learning model. We used the k-Fold Cross-Validation procedure (resampling). The CV (k-Fold Cross-Validation) is donated with the following equation:

$$CV(\hat{f}) = \frac{1}{N}\sum_{i=1}^{N} L\left(y_i, \hat{f}^{-k(i)}\left(x_j\right)\right) \qquad (5)$$

where i is the observation by randomization

$$\hat{f}^{-k}(x)$$

is the fitted function which is computed with the kth part of the data removed. K = N (leave-one-out) cross-validation, ad k(i) = i for the ith observation and the fit is computed using all data except the ith. Typically K choices are 5 or 10.

i: CONFIDENCE INTERVAL

The confidence interval quantifies the uncertainty of an estimate for the predicted outcomes of the evaluated classifiers. Therefore, the confidence interval is achieved [50] by calculating the formula:

$$CI = \bar{x} \pm z\frac{s}{\sqrt{n}} \qquad (6)$$

where, CI = confidence interval $\bar{x}$ = sample mean z = confidence level value $\boldsymbol{s}$ = sample standard deviation n = sample size

j: LEFT

We have also used the measures Log Loss [51] and the Left-Curves (equation 9) for the models' predictions' explainability stage:

$$\text{Left } = \frac{\text{Predicted Rate}}{\text{Average Rate}} \qquad (7)$$

III. RESULTS

This section reports the proposed ICU predictive framework results from a real-hospital dataset (Al-Ain Hospital). We evaluate the predictive framework using Cross-Validation (stage 1) and with two portions (training and testing) to put the model into the practical aspect (stage 2). We explain the LOS predicted results within the LOS framework for the best performing model (stage 3).

A. CROSS-VALIDATION RESULTS (STAGE 1: BASELINING)

In this work, a k-Fold (k = 5) Cross-Validation was utilized (k groups approach). It is a practical procedure, especially to evaluate the classifier's performance when we have limited data. Hyperparameters optimization is used to assess the predictive framework using k-fold with the best model's tuned parameters. Experiment 1 in Table 2 reports the best results in terms of (Accuracy, Precision Sensitivity, Specificity and AUC) amongst all five experiments. In experiment 1, the XGboost model showed relatively better results than other

Algorithm 2 Outperforming Models Evaluation Stage With Hyperparameters

```
Require:
    $EHRfeatures, f_c = [fc_1, fc_2, \ldots, fc_n]$;
    Dataset; Target label: Short, OS, Long OS;
    $MP$: Outperformed models from Stage 1 with hyperparameters.
    $Winning\_best\_modelMB$: Evaluated models on training matrix and testing matrix.
Ensure: Split $EHR_{features}$ into $DTR$ (training matrix) and $DTS$ (testing matrix)
0: Initialize $MTR$ and $MTS$ as empty arrays.
0: for each $m$ in $MP$ do
0:     Train model $m$ using $DTR$.
0:     Compute model performance $AUC$ for $m$ on $MTR$.
0: end for
0: for each $n$ in $MP$ do
0:     Train model $n$ using $DTS$.
0:     Compute model performance $AUC$ for $n$ on $MTS$.
0: end for
0: $MB \leftarrow$ Model from $MP$ with highest $AUC$ on $MTR$ and $MTS$.
0: return Winning best model $MB$.=0
```

models (RF, GB, MLP, and LR), particularly using the AUC model's measured outcomes (77.9% and 74.4%) for validation (V) and testing (T) scores. In comparison, MLP recorded an AUC of (59.4% and 57.5%) for V and T, respectively.

Hence, experiment 1 produced the best predictive LOS performance compared to the other experiments in Table 2. We have selected Experiment 1 for a further look at the estimated skill of a classification method using the calculated confidence interval of 95% (95%CI). Table 3 indicates the classification error of each model with metrics (Accuracy, Precision Sensitivity, Specificity and AUC) with the CI of 95% or the true classification error of each model is likely to be within the range of the +/- CI 95% values.

The calculated average overall models' performance of all experiments in Table 2 within scenario 1 revealed a preference towards the XGboost Model. Therefore, XGboost is evidenced to be a robust classifier. Hence, XGBoost is our selection in the baselining stage (Fig 1).

B. OUTPERFORMING MODELS EVALUATION (STAGE 2: PRACTICAL EXPERIMENTATION)

In this stage, we experimented with the five main experiments (G+C+L+M, G, C, L, M) with the same tuned models' parameters are used in the previous stage (baselining). For this purpose, and based on the previous stage (baselining), we pick the most outperforming models for a further performance evaluation on a practical aspect to attest to the candidate models' performance. The result of this experimental procedure is the best performing model. The performance measure is the ROC curves that display TPR on Y-axis and FPR on the X-axis. For the practical experimentation step, we have used two portions: the training set (66%) and the testing (34%). The performances of the three models in the practical experimentation stage were close to each other. For example, when evaluating the ROC for the XGBoost classifier, we can observe that the set of all of the features experiments (G+C+L+M) achieved the best ROC (88%: CI%95 [81.6%-94.4%]) for short LOS and Long LOS. The second apparent results are with the Medication (M) ICU features, where the XGBoost achieved (ROC= 84%: [76.8%-91.2%]) for Short LOS and Long LOS classes. This confirms the model's ability to commence fewer prediction errors in both classes (short and long LOS). The Gradient Boosting achieved comparatively comparable results to the XGboost in all features experiment (G+C+L+M), with ROC (88%: CI%95 [81.6%-94.4%]) for short LOS and Long LOS. However, Gradient Boosting obtained slightly better results with Medication features (M) experiments, achieving ROC (85%: CI%95[78%-92%]) for Short LOS and Long LOS, respectively. At the same time, the XGboost results were slightly better than GB in the general (G) and clinical (C) experiments. However, the General ROC results attained (XGboost 57% and Gradient Boosting 55%), Clinical Short and Long LOS managed to obtain (52%: CI%95[42.2%-61.8%], and 51% CI%95[41.2%-60.8%]) for XGBoost, and Gradient Boosting respectively as the least important ROC results in the experimentation stage. Moreover, XGboost attained ROC of (64%: CI%95[54.6%-73.4%]) and (62%: CI%95[52.5%-71.5%]) for GB in the laboratory experiment. Finally, the Random forest achieved steady ROC results for both classes (Short and Long) LOS. In all features (G+C+L+M), the RF obtained the highest ROC (88%: CI%95 [81.6%-94.4%]) within the five experiments, then Medication experiments with ROC of (85%: CI%95[78%-92%]), and (64%: CI%95[54.6%-73.4%]) for laboratory, (54%: CI%95[44.2%-63.8%]) (57%: CI%95[47.3%-66.7%]) for clinical and general experiments respectively.

C. EXPLAINING THE XGBOOST PREDICTIVE RESULTS (STAGE 3)

This section explains the predictive results of the winning model in the proposed framework. The winning model is the most robust classifier based on the predicted outcomes and ability to attain stable and reliable results based on different experimentation setups from Al-Ain hospital data. The XGboost achieved the desired outcomes; therefore, it is our selection model for further result explanation. We aim to reveal the black box of the predictive classification model (XGboost) and make it more understandable and easy to explain for non-machine learning people. This may include healthcare workers in hospitals and healthcare givers such as hospital managers, clinicians, hospital nurses, and health insurance companies. The predictive outcomes are explained from two perspectives. The first approach is the classification outcomes (overall) classifier explainability using the whole

TABLE 2. A comparison between features selection sets based on the patient's information profile (Cross-validation and testing reported results). LOS predictive framework on a real hospital data (Al-Ain hospital, UAE).

height Model	Accuracy %		Precision %		Sensitivity %		Specificity %		AUC %	
	V	**T****	**V**	**T**	**V**	**T**	**V**	**T**		**T**
Experiments 1: All dataset ($\mathbf{G}+\mathbf{C}+\mathrm{L}+\mathrm{M}$) ICU features										
XGBoost	83.7	78.9	87.3	83.2	90.8	86.7	64.9	62.1	77.9	74.4
RF	82.3	77	83.6	78.8	94.1	90.9	50.8	47	72.5	68.9
GB	80.9	74.6	84.6	81.2	90.1	81.8	56.1	59	73.1	70.5
MLP	73.7	70.3	77.1	72.1	90.8	92.3	28.1	22.7	59.4	57.5
LR	77.5	69.9	84.8	80.8	84.2	73.4	59.6	62.1	71.9	67.8
Experiments 2: General (G) ICU features										
XGBoost	72.2	68.9	73.3	69.5	97.4	97.2	5.2	7.5	51.3	52.4
RF	70.8	67	73.8	69.3	92.8	93	12	10.6	52.5	51.8
GB	70.8	67.9	72.9	69.2	95.4	95.8	5.2	7.5	50.3	51.7
MLP	72.7	69.4	73.4	69.7	98	97.9	5.2	7.5	51.6	52.7
LR	72.7	69.4	73.2	69.5	98.7	98.6	3.5	7.5	51.1	52.3
Experiments 3: Clinical (C) ICU features										
XGBoost	72.7	68.4	72.7	68.4	100	100	0	0	50	50
RF	73.2	68.4	73.3	68.8	99.3	98.6	3.5	3	51.4	50.8
GB	72.7	68.4	72.7	68.4	100	100	0	0	50	50
MLP	72.7	67.5	72.7	68.1	100	98.6	0	0	50	49.3
LR	72.7	67.5	72.7	68.4	100	100	0	0	50	50
Experiments 4: laboratory (L) ICU features										
XGBoost	71.8	67.5	72.9	68.5	97.4	97.2	3.5	3	50.4	50.1
RF	72.7	68.4	72.9	68.4	99.3	100	1.7	0	50.5	50
GB	73.2	67.9	73.3	68.1	99.3	98.6	3.5	0	51.4	49.3
MLP	72.2	67.9	73	68.3	98	99.3	3.5	0	50.8	49.7
LR	72.2	67.9	73	68.3	98	99.3	3.5	0	50.8	49.7
Experiments 5: Medication (M) ICU features										
XGBoost	79.9	77	85.3	79.9	87.5	88.8	59.6	51.5	73.6	70.2
RF	77	76.6	81	95.1	89.5	95.1	43.8	36.3	66.7	65.7
GB	82.3	75.1	86.6	89.5	89.5	89.5	63.1	43.9	76.3	66.7
MLP	80.9	78.5	85	89.5	89.5	89.5	57.8	54.5	73.7	72
LR	81.8	78	87	89.5	88.2	89.5	64.9	53	76.5	71.3
Models' average performance of all experiments										
XGBoost	76.06	72.14	78.3	73.9	94.62	93.98	26.64	24.82	60.64	59.42
RF	75.2	71.48	76.92	76.08	95	95.52	22.36	19.38	58.72	57.44
GB	75.98	70.78	78.02	75.28	94.86	93.14	25.58	22.08	60.22	57.64
MLP	74.44	70.72	76.24	73.54	95.26	95.52	18.92	16.94	57.1	56.24
LR	75.38	70.54	78.14	75.3	93.82	92.16	26.3	24.52	60.06	58.22

TABLE 3. Baselining predictive models on Al-Ain dataset (Experiments 1, "G+C+L+M") with hyperparameters and cross-validation (K-fold = 5) approaches & 95%CI.

Model	Accuracy % + 95% CI		Precision % + 95% CI		Sensitivity % + 95% CI		Specificity % + 95% CI		AUC % + 95% CI	
	V*	T**	V	T	V	T	V	T	V	T
XGBoost	83.7 (76.5-90.9)	78.9 (70.9-86.9)	87.3 (80.8- 93.8)	83.2 (75.9 – 90.5)	90.8 (85.1 – 96.5)	86.7 (80- 93.4)	64.9 (55.5 – 74.3)	62.12 (52.6- 71.6)	77.9 (69.8 - 86)	74.4 (65.8 - 83)
RF	82.3 (74.8-89.8)	77 (68.8-85.2)	83.6 (76.3-90.9)	78.8 (70.8 – 86.8)	94.1 (89.5 – 98.7)	90.9 (85.3 – 96.5)	50.8 (41-60.6)	46.96 (37.2-56.7)	72.5 (63.7-81.3)	68.9 (59.8-78)
GB	80.9 (73.2-88.6)	74.6 (66.1-83.1)	84.6 (77.5-91.7)	81.2 (73.5-88.9)	90.1 (84.2-96)	81.8 (74.2-89.4)	56.1 (46.4-65.8)	59 (49.4-68.6)	73.1 (64.4-81.8)	70.5 (61.6-79.4)
MLP	73.7 (65.1-82.3)	70.3 (61.3-79.3)	77.1 (68.9-85.3)	72.1 (63.3-80.9)	90.8 (85.1-96.5)	92.3 (87.1-97.5)	28.07 (19.3-36.9)	22.72 (14.5-30.9)	59.4 (49.8-69)	57.5 (47.8-67.2)
LR	77.5 (69.3-85.7)	69.9 (60.9-78.9)	84.8 (77.8-91.8)	80.8 (73.1-88.5)	84.2 (77.1-91.3)	73.4 (64.7-82.1)	59.64 (50-69.3	62.1 (52.6-71.6)	71.9 (63.1-80.7)	67.8 (58.6-77)

dataset. We refer to them as (predictive outcomes with the model's overall explainability) in the ICU dataset. The second approach is where we put the patients in the perception of the explanation. We refer to it as model's patient-centred prediction outcome explainability. For this purpose, we have exploited the ExplainerDashboard prediction explainer [52] as the explainable artificial intelligence (XAI) tool. The XAI tool builds explainable interactive dashboards to analyze the classification and prediction results. ExplainerDashboard XAI libraries are compatible with Python. The dashboard is running on the local server of the experimenting computing instance. The XGboost classifier is used with hyperparameters values per table (Table 1).

1) PREDICTIVE OUTCOMES OF XGBOOST OVERALL CLASSIFIER EXPLAINABILITY

a: XGBOOST PERFORMANCE METRICS EXPLAINABILITY

The XGboost model's performance metrics for Short and Long Length of Stay (LOS) labels are reported in Table 4. The achieved accuracy for both classes is 94.6%, with precision of

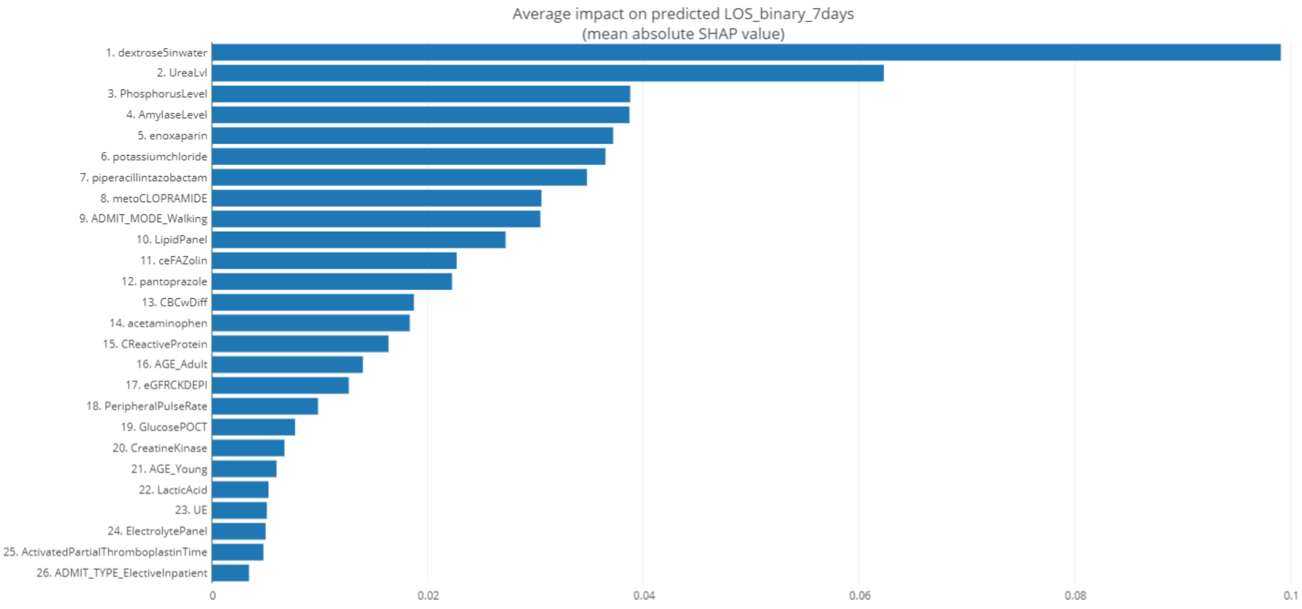


FIGURE 2. Features importance by mean absolute SHAP value: XGboost.

TABLE 4. Model performance metrics for short LOS vs. Long LOS for XGBoost classifier: the XAI tool.

Metric	Short LOS	Long LOS
Accuracy	94.6%	94.6%
Precision	91.4%	96%
Recall	90.6%	96.4%
F1-Score	91%	96.2%
ROC-AUC_	98%	98%
PR-AUC_	95.6%	99.1%
Log Loss	0.211	0.211

91.4% and 96% for Short and Long LOS labels, respectively. XGboost demonstrated robustness in differentiating between the predicted Short and Long LOS classes, with a receiver operating characteristic area under curve (ROC-AUC) of 98%. The model's recall and F1-Score results were also relatively close. These outcomes confirm XGboost's ability to attain desired predictions with negligible error rates (2%).

Fig 2 illustrates the average impact of features on the predicted LOS labels (Short and Long) based on the mean absolute SHAP value. Medication and laboratory information contributed significantly to the XGboost model's decision, followed by general admission features, such as admission mode (walking) and age (Young Adult).

Notably, clinical features did not appear in the features importance plot. This could be due to the SHAP method, which measures a feature's global influence by comparing model predictions with and without the feature. The SHAP values provide information about each feature's contribution to individual predictions. Fig 3 shows the classification plot of the Short and Long LOS labels. The cutoff for the classification report is set at 80% per the XAI method. Of the XGBoost classified Short LOS cases, 91.43% are above the cutoff (Fig 3.a), while 96% of the XGboost Long LOS cases are above the cutoff (Fig 3.b). These findings demonstrate XGboost's effectiveness in differentiating between the two labels.

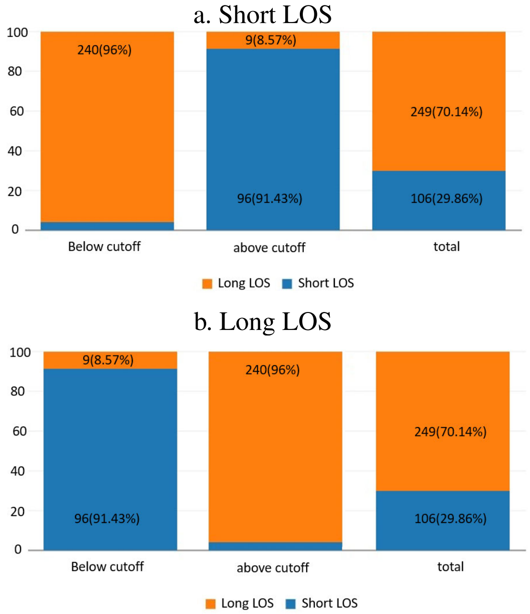
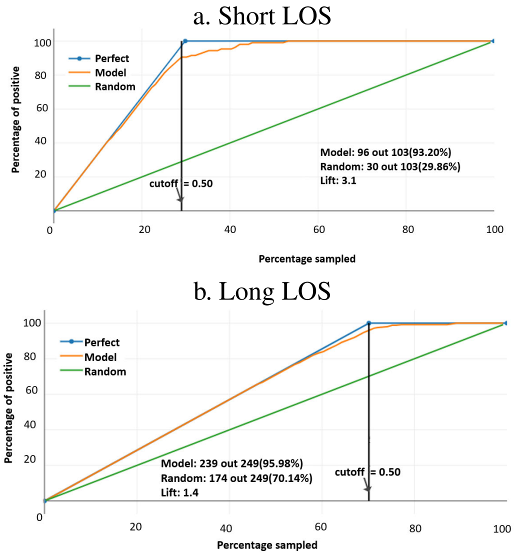


FIGURE 3. Percentage above and below cutoff(Short LOS & Long LOS).

The left curve (Fig 4) is an important measure that helps with the predictive classification model's effectiveness (XGboost). For any given number of cases (percentage of samples: Fig 4), it illustrates the expected number of positives we would predict if we did not have a model but simply selected random cases. The left curve provides a benchmark



a. Short LOS

b. Long LOS

FIGURE 4. Left Curves for the XGBoost classification results (Short LOS & Long LOS).

against which we can see the model's performance [54]. For example, based on equation (9), the XGBoost model gives us a left 3.1 for predicting the Short LOS class and a lift of 1.4 for predicting the Long LOS class. A good classifier will provide us with a high lift when we act on only a few cases, and as we include more cases, the lift will decrease. The left curve with the best classifier (that commences fewer errors) would overlap with the existing curve at the start, then continue with a slope of 1 until it (all successes), then continue horizontally to the right [54]. This is clearly projected in Fig 4 (a & b). Thus, left curves aid beds' managers or healthcare decision-makers to understand the decision made by the classification model and how it impacts the healthcare decisions and strategies for managing resources utilization and beds availability. The predictive outcomes overall classifier explainability approach is clearly proven that the XGboost to be robust with high and desired predictive outcomes (Short and Long LOS). Furthermore, the XAI tool explained the XGBoost classifier's inner workings with ease and made the decision predicted outcomes clear and understandable to a non-data machine learning specialist.

2) PATIENT-CENTERED PREDICTION OUTCOMES EXPLAINABILITY

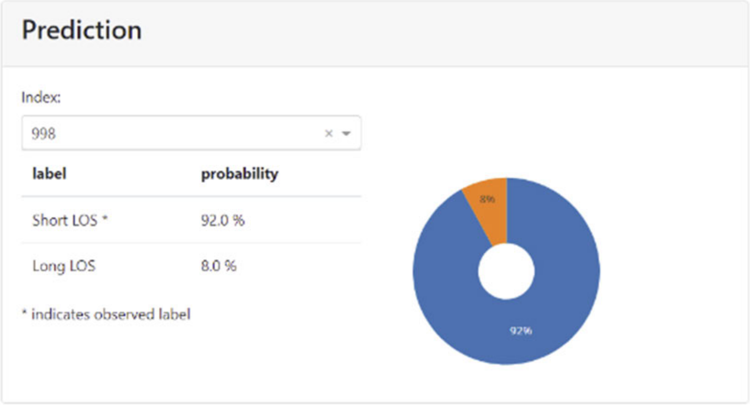
We assess the ability and the inner workings to explain the decision of the XGBoost. We selected a random patient (de-identified) with the given data index (998). The index values are the numerical order in the Pandas DataFrame in Python. Therefore, the number 998 represents the patient's case (patient profile) or admitted information (General, Clinical, Laboratory and Medication) in the dataset. The XAI tool provides a range of explainable prediction components, including individual prediction explainability. We have utilized the individual prediction components, the (pdp) plot of the feature, the contribution to prediction probability, the dependents plot of the feature, and the prediction percentage of each class label. In addition, we used the "UreaLvl" feature to evaluate the XGboost performance at a feature and an individual patient level. The selection of the UreaLvl is indiscriminate and only to explain XGboost evaluation at the patient-centred level.

We evaluate each single feature in Fig 5 and how it affects the XGboost prediction. The contribution of features to the model's outcomes shows the breakdown of every single feature in the XGBoost and how it affected the final prediction for patient 998. The breakdown shows how the model thinks that patient 998 was predicted to be a short stay.

Fig 6 shows that the final contribution to the XGBoost prediction is (91.96%), and this is because the features contributed have a high predicted probability in the patient case. This is justified by the electronic health records (Al-Ain hospital) features present in the patient's case. Therefore, the more (General, Clinical, Laboratory and Medication) features we input into the model, the more ability the model has to provide a reliable prediction. The features per Fig 6 (a & ') represent how each feature adds up to the final contribution of the Short LOS prediction. Partial Dependent Plot (pdp) Fig 7 shows how the prediction changes based on each feature input. For instance, Fig 7 (a) clearly shows the partial feature contribution to the prediction of Short LOS outcomes, but if we look at the same feature from the perspective of the Long LOS label, we can see a weak partial contribution to the predicted outcomes. Fig 8 (a & b) epitomizes the dependence plot of UreaLvl according to SHAP values in relation to dextrose5inwater. We can see the relationship (random selection) between "UreaLvl and dextrose5inwater" features and their impact on the XGBoost prediction outcomes (Short and Long LOS). Finally, the XAI tool provides an overall picture of the XGBoost predicted outcomes or the prediction decision of each class label at the patient level (Fig 5). For example, the XGboost results designate that the patient is likely to stay (Short LOS) with a probability of 92% and likelihood of not staying (Long LOS) with 8%. Eventually, and after explaining the inner workings of the XGboost from the features perspective and their effect on the models' prediction outcomes. Also, the interaction between two features contributes to the overall picture of the final prediction, the XAI tool provides the ability for us to understand the inner details of the chosen or the winning and outperforming model (XGBoost). Notably, these explanations may guide the concerned person of the AI model, such as the bed manager, clinicians, and healthcare insurance companies, to investigate each feature or features interactions on the model's outcomes or find unexplored relationships between interacted features and their interactions impacting on the model's prediction classes.

IV. DISCUSSION

One of the most anticipated tasks in this study is to evaluate the suggested LOS framework utilising cross-validation and practical experimentation approaches with hyperparameters



Prediction

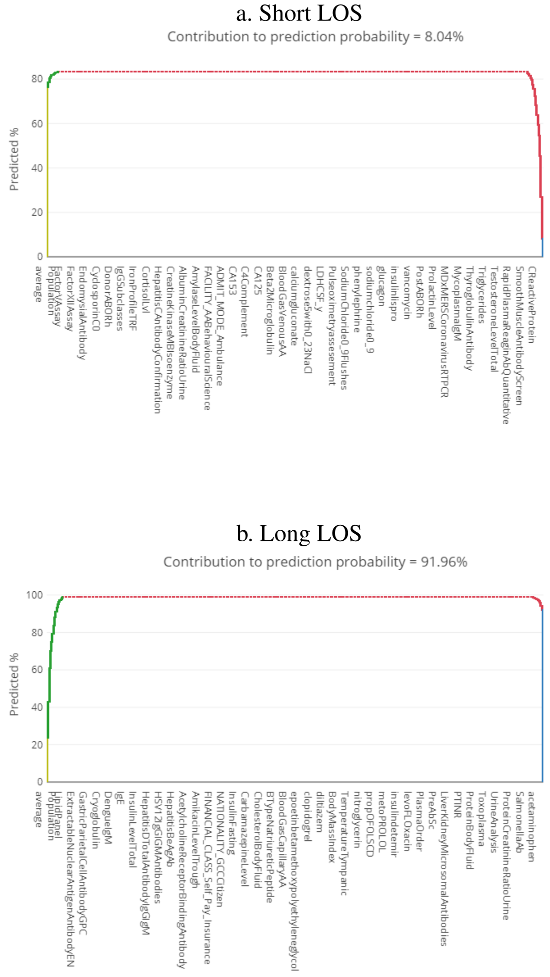
Index:

998

label	probability
Short LOS *	92.0 %
Long LOS	8.0 %

* indicates observed label

FIGURE 5. Prediction percentage by class labels (Short LOS and Long LOS).



a. Short LOS

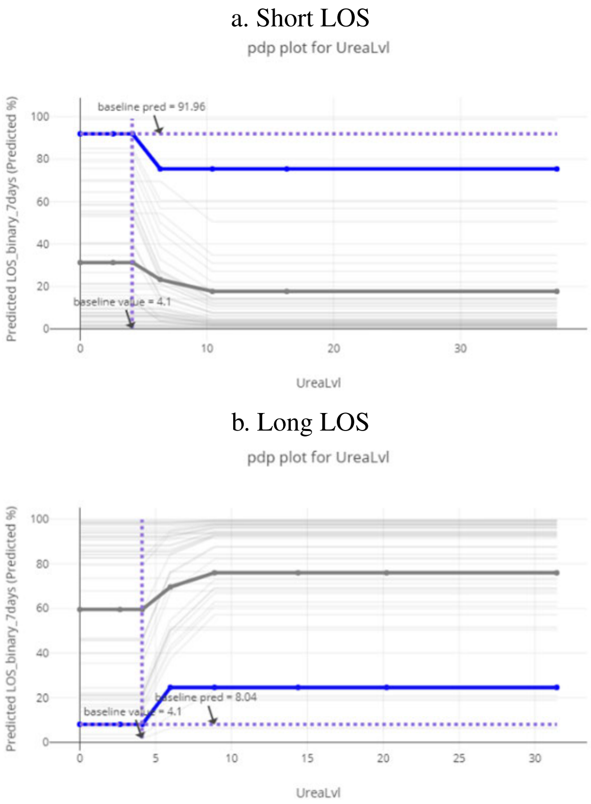
Contribution to prediction probability = 8.04%

b. Long LOS

Contribution to prediction probability = 91.96%

FIGURE 6. Contribution to the prediction probability (Short and Long) LOS classes.

tuning. The ICU-LOS prediction framework produced low variance and low bias with both approaches in all studies (stage 1). This validates the durability of using validation in a cross-validation technique to determine the optimum parameters of each model (model tuning) and produce improved predicted outcomes. Furthermore, given these prevalent challenges in the machine learning area, this technique can assist to decrease overfitting and prevent underfitting. This eventually leads to better predicted machine learning performance.



a. Short LOS

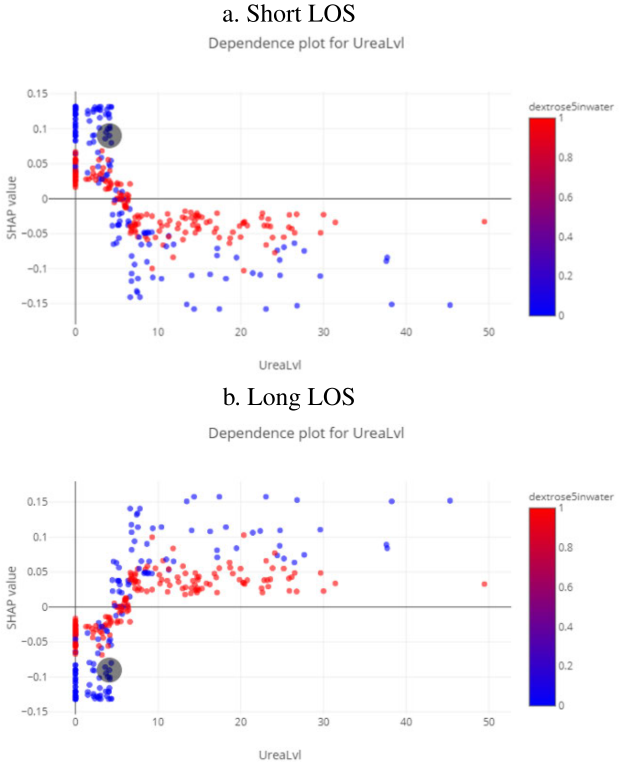
pdp plot for UreaLvl

b. Long LOS

pdp plot for UreaLvl

FIGURE 7. Partial dependent plot (pdp) for UreaLvL feature (Short and Long) LOS classes.

During the cross-validation (baselining) stage of model performance evaluation, we applied statistical inference (Confidence Intervals "CI 95%"). The CI%95 provides interval estimators for the prediction error and is a more useful


a. Short LOS

b. Long LOS

FIGURE 8. Dependence plot for UreaLvl: (Short and Long) LOS classes.

means of analysing and interpreting predictive outcomes. As a result, the uncertainty estimate assists hospital bed management in determining how well or poorly the LOS classifiers function. Furthermore, it is beneficial to consider which model is less complex and more interpretable. The trials (stage 1) demonstrated that the proposed LOS predictive framework could be a feasible method. For example, the AUC of the XGBoost validation set is 77.9% (69.8% - 86%) with CI%95, while the AUC of the testing set is 74.4% (65.8% - 83%). This indicates that the XGboost can distinguish between the two labels (Short LOS and Long LOS) on the validation set with an AUC of 77.9%, while the true classification error of the XGboost is likely between 69.8% and 86%. Similarly, the AUC of the XGboost Model on the testing set is 74.4%, and it is likely to correctly forecast the LOS labels within 65.8%- 83%. Thus, the cross-validation with hyperparameter optimization strategy (stage 1) was advantageous and yielded the expected results for predicting short and long LOS with low variation and bias.

The validation procedure consists of evaluating the model fit on the training dataset while tuning the model hyperparameters. The second approach to model evaluation tests the robustness of models during practical application in real-world scenarios. To ensure reliable predictions for short and long LOS in newly admitted ICU patients, it is essential to validate the performance of the selected predictive model before practical use. The predictive performance of the three candidate models was reasonably comparable, with a slight preference for XGBoost, which produced superior results when using all dataset features (G+C+L+M). The performance of the models improved as additional features were added. The bagging and boosting approaches are frequently used in ensemble learning as they generate accurate and robust models that are suitable for two-class classification problems.

The proposed LOS architecture was chosen based on classifiers' performance. Thus, we used XGboost to explain artificial intelligence to non-specialists like hospital bed administrators, healthcare workers, and CDSS workers. In predictive classification tasks, this is referred to as opening the black box (model explainability). This is critical for clinical decision support systems (CDSS) help clinicians make informed decisions and predict inpatient health outcomes [55]. The requirements for CIS systems go beyond the performance of the model [56]. The CDSSs are established in clinical settings to exhibit proven safety [57]. Therefore, ensuring the interpretability and transparency of machine learning (ML) models is crucial for their safe and effective use in CDSS. To address this, we employed an XAI tool to explain the predictions of our XGboost classifier and provide insight into its inner workings.

This approach not only helps hospital decision-makers and bed managers to understand the reasoning behind the predictions but also enables non-ML professionals to evaluate the model's operational performance. At the micro-level, it is essential to explain each patient's prediction, enabling healthcare workers to make informed decisions about patient care and resource allocation. The XAI tool used here provides a crucial indicator of whether individuals understand the model's limitations, allowing them to identify any missing information and overrule the model if necessary. By enhancing transparency and interpretability, XAI tools can improve the trustworthiness and usefulness of ML models in healthcare decision-making.

The XGBoost classifier exhibited strong predictive ability, with Accuracy, Precision, F1-score, Recall, ROCAUC, and PR-AUC values for both Short and Long LOS greater than 90%. The model's explainability through XAI techniques enables healthcare workers to identify which patient characteristics contribute to Short or Long LOS. This information helps to allocate resources effectively and aids in the development of safe and trustworthy clinical decision support systems. The machine learning model's capacity to provide data-driven predictions supports evidence-based decision-making in healthcare.

The study's primary beneficiaries are hospital healthcare workers, including bed managers, ICU clinicians, and nurses, who can benefit from a comprehensive and explainable predictive framework for predicting inpatient length of stay at ICU admission or transfer. The framework provides advanced patient health monitoring capabilities by leveraging artificial intelligence advancements, enabling healthcare personnel to make better judgments in a dynamic and demanding health environment. The ICU-LOS framework can increase patient flow in the ICU, optimize resource capacity (hospital beds, ventilators), and improve hospital resource allocation. Effective bed management and resource management can reduce healthcare spending, waste time, and improve service quality. Incorporating the architecture into hospital CIS systems can enhance productivity and operational efficiency. Poor hospital administration can lead to overburdened healthcare workers, an overabundance of patients, and a drain on hospital resources, resulting in rejected admissions. Therefore, implementing the presented framework can significantly benefit hospital healthcare professionals, health care providers, and healthcare stakeholders in attaining the critical goal of accommodating newly admitted patients, even during uncertain times.

V. CONCLUSION AND FUTURE WORK

This study developed a predictive ICU framework using real hospital data to predict patients' length of stay at ICU admission. This practical framework offers significant implications for ICU bed management and resource utilization, achieving desired predictive results through its three-stage LOS predictive process. Among the various models tested, the XGBoost model emerged as the best performer due to its ability to provide explainable results to non-AI professionals. Notably, this study is the first to present an AI-explainable framework for predicting ICU patients' length of stay using a data-driven approach. The proposed framework is versatile, applicable across various diseases and health conditions, making it valuable for clinical research and electronic health records. It also has the potential to improve predictive tasks such as identifying patients at risk of mortality. Future research will focus on integrating user-centered clinical predictive systems into daily hospital workflows and thoroughly investigating the use of explainable AI in hospital, emergency department, and ICU settings. This will help establish genuine ML-xAI implementation and standardize their use in electronic health records and healthcare systems.

REFERENCES

[1] A. Awad, M. Bader-El-Den, and J. McNicholas, "Patient length of stay and mortality prediction: A survey," *Health Services Manage. Res.*, vol. 30, no. 2, pp. 105–120, May 2017.

[2] OECD. (2020). *Length of Hospital Stay (Indicator)*. Accessed: Jul. 21, 2021. [Online]. Available: https://data.oecd.org/healthcare/length-of-hospital-stay.htm

[3] Australian Institute of Health and Welfare, Canberra, ACT, Australia. (2011). *Hospital Performance: Length of Stay in Public Hospitals in 2011–12*. [Online]. Available: https://www.aihw.gov.au/reports/hospitals/hospital-performance-length-of-stay-in-2011-12

[4] F. Pecoraro, F. Clemente, and D. Luzi, "The efficiency in the ordinary hospital bed management in Italy: An in-depth analysis of intensive care unit in the areas affected by COVID-19 before the outbreak," *PLoS ONE*, vol. 15, no. 9, Sep. 2020, Art. no. e0239249.

[5] M. Hassan, H. P. Tuckman, R. H. Patrick, D. S. Kountz, and J. L. Kohn, "Hospital length of stay and probability of acquiring infection," *Int. J. Pharmaceutical Healthcare Marketing*, vol. 4, no. 4, pp. 324–338, Nov. 2010.

[6] M. C. Blom, K. Erwander, L. Gustafsson, M. Landin-Olsson, F. Jonsson, and K. Ivarsson, "The probability of readmission within 30 days of hospital discharge is positively associated with inpatient bed occupancy at discharge—A retrospective cohort study," *BMC Emergency Med.*, vol. 15, no. 1, pp. 1–6, Dec. 2015.

[7] E. Rocheteau, P. Liò, and S. Hyland, "Temporal pointwise convolutional networks for length of stay prediction in the intensive care unit," 2020, *arXiv:2007.09483*.

[8] C. W. Hanson, C. S. Deutschman, H. L. Anderson, P. M. Reilly, E. C. Behringer, C. W. Schwab, and J. Price, "Effects of an organized critical care service on outcomes and resource utilization: A cohort study," *Crit. Care Med.*, vol. 27, no. 2, pp. 270–274, Feb. 1999.

[9] S. Siddiqui, S. Ahmed, and R. Manasia, "Apache II score as a predictor of length of stay and outcome in our ICUs," *J. Pakistan Med. Assoc.*, vol. 55, no. 6, p. 253, 2005.

[10] W. A. Knaus, J. E. Zimmerman, D. P. Wagner, E. A. Draper, and D. E. Lawrence, "APACHE—Acute physiology and chronic health evaluation: A physiologically based classification system," *Crit. Care Med.*, vol. 9, no. 8, pp. 591–597, Aug. 1981.

[11] A. Knaus, D. P. Wagner, E. A. Draper, J. E. Zimmerman, M. Bergner, P. G. Bastos, C. A. Sirio, D. J. Murphy, T. Lotring, A. Damiano, and F. E. Harrell Jr., "The APACHE III prognostic system: Risk prediction of hospital mortality for critically III hospitalized adults," *Chest*, vol. 100, no. 6, pp. 1619–1636, 1991.

[12] M. T. Keegan, O. Gajic, and B. Afessa, "Comparison of APACHE III, APACHE IV, SAPS 3, and MPM0III and influence of resuscitation status on model performance," *Chest*, vol. 142, no. 4, pp. 851–858, Oct. 2012.

[13] C.-C. Yeh, Y.-A. Chen, C.-C. Hsu, J.-H. Chen, W.-L. Chen, C.-C. Huang, and J.-Y. Chung, "Quick-SOFA score ≥ 2 predicts prolonged hospital stay in geriatric patients with influenza infection," *Amer. J. Emergency Med.*, vol. 38, no. 4, pp. 780–784, Apr. 2020.

[14] C. Li, L. Chen, J. Feng, D. Wu, Z. Wang, J. Liu, and W. Xu, "Prediction of length of stay on the intensive care unit based on least absolute shrinkage and selection operator," *IEEE Access*, vol. 7, pp. 110710–110721, 2019.

[15] M. M. Islam, T. N. Poly, and Y.-C. Li, "Recent advancement of clinical information systems: Opportunities and challenges," *Yearbook Med. Informat.*, vol. 27, no. 1, pp. 83–90, Aug. 2018.

[16] E. Levesque, E. Hoti, D. Azoulay, P. Ichai, D. Samuel, and F. Saliba, "The implementation of an intensive care information system allows shortening the ICU length of stay," *J. Clin. Monitor. Comput.*, vol. 29, no. 2, pp. 263–269, Apr. 2015.

[17] S. Calloway, H. A. Akilo, and K. Bierman, "Impact of a clinical decision support system on pharmacy clinical interventions, documentation efforts, and costs," *Hospital Pharmacy*, vol. 48, no. 9, pp. 744–752, Sep. 2013.

[18] S. T. McMullin, T. P. Lonergan, C. S. Rynearson, T. D. Doerr, P. A. Veregge, and E. S. Scanlan, "Impact of an evidence-based computerized decision support system on primary care prescription costs," *Ann. Family Med.*, vol. 2, no. 5, pp. 494–498, Sep. 2004.

[19] R. T. Sutton, D. Pincock, D. C. Baumgart, D. C. Sadowski, R. N. Fedorak, and K. I. Kroeker, "An overview of clinical decision support systems: Benefits, risks, and strategies for success," *NPJ Digit. Med.*, vol. 3, no. 1, pp. 1–10, Feb. 2020.

[20] D. Romanow, A. Rai, M. Keil, and S. Luxenberg, "Does extended CPOE use reduce patient length of stay?" *Int. J. Med. Informat.*, vol. 97, pp. 128–138, Jan. 2017.

[21] G. S. Birkhead, M. Klompas, and N. R. Shah, "Uses of electronic health records for public health surveillance to advance public health," *Annu. Rev. Public Health*, vol. 36, no. 1, pp. 345–359, Mar. 2015.

[22] B. Shickel, P. J. Tighe, A. Bihorac, and P. Rashidi, "Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis," *IEEE J. Biomed. Health Informat.*, vol. 22, no. 5, pp. 1589–1604, Sep. 2018.

[23] C. Combi and G. Pozzi, "Clinical information systems and artificial intelligence: Recent research trends," *Yearbook Med. Informat.*, vol. 28, no. 1, pp. 83–94, Aug. 2019.

[24] E. Berner and T. L. Lande, *Clinical Decision Support Systems* (Heath Informatics). Springer, 2007. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-31913-1_1

[25] X. Ma, Y. Si, Z. Wang, and Y. Wang, "Length of stay prediction for ICU patients using individualized single classification algorithm," *Comput. Methods Programs Biomed.*, vol. 186, Apr. 2020, Art. no. 105224.

[26] L. Su, Z. Xu, F. Chang, Y. Ma, S. Liu, H. Jiang, H. Wang, D. Li, H. Chen, X. Zhou, N. Hong, W. Zhu, and Y. Long, "Early prediction of mortality, severity, and length of stay in the intensive care unit of sepsis patients based on sepsis 3.0 by machine learning models," *Frontiers Med.*, vol. 8, p. 883, Jun. 2021.

[27] P. V. Staziaki, D. Wu, J. C. Rayan, I. D. D. O. Santo, F. Nan, A. Maybury, N. Gangasani, I. Benador, V. Saligrama, J. Scalera, and S. W. Anderson, "Machine learning combining CT findings and clinical parameters improves prediction of length of stay and ICU admission in torso trauma," *Eur. Radiol.*, vol. 31, no. 7, pp. 5434–5441, Jul. 2021.

[28] K. Alghatani, N. Ammar, A. Rezgui, and A. Shaban-Nejad, "Predicting intensive care unit length of stay and mortality using patient vital signs: Machine learning model development and validation," *JMIR Med. Informat.*, vol. 9, no. 5, May 2021, Art. no. e21347.

[29] A. E. W. Johnson, T. J. Pollard, L. Shen, L.-W.-H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark, "MIMIC-III, a freely accessible critical care database," *Sci. Data*, vol. 3, no. 1, pp. 1–9, May 2016.

[30] T. Gentimis, A. J. Alnaser, A. Durante, K. Cook, and R. Steele, "Predicting hospital length of stay using neural networks on MIMIC III data," in *Proc. IEEE 15th Int. Conf Dependable, Autonomic Secure Comput., 15th Int. Conf Pervasive Intell. Comput., 3rd Int. Conf Big Data Intell. Comput. Cyber Sci. Technol. Congr. (DASC/PiCom/DataCom/CyberSciTech)*, Nov. 2017, pp. 1194–1201.

[31] R. J. Steele and B. Thompson, "Data mining for generalizable preadmission prediction of elective length of stay," in *Proc. IEEE 9th Annu. Comput. Commun. Workshop Conf. (CCWC)*, Jan. 2019, pp. 127–133.

[32] H. Quan, V. Sundararajan, P. Halfon, A. Fong, B. Burnand, J.-C. Luthi, L. D. Saunders, C. A. Beck, T. E. Feasby, and W. A. Ghali, "Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data," *Med. Care*, vol. 43, no. 11, pp. 1130–1139, Nov. 2005.

[33] M. Maniruzzaman, M. J. Rahman, M. Al-MehediHasan, H. S. Suri, M. M. Abedin, A. El-Baz, and J. S. Suri, "Accurate diabetes risk stratification using machine learning: Role of missing value and outliers," *J. Med. Syst.*, vol. 42, no. 5, pp. 1–17, May 2018.

[34] B. Alsinglawi, F. Alnajjar, O. Mubin, M. Novoa, O. Karajeh, and O. Darwish, "Benchmarking predictive models in electronic health records: Sepsis length of stay prediction," in *Proc. 34th Int. Conf. Adv. Inf. Netw. Appl.*, 2020, pp. 258–267.

[35] B. Alsinglawi, F. Alnajjar, O. Mubin, M. Novoa, M. Alorjani, O. Karajeh, and O. Darwish, "Predicting length of stay for cardiovascular hospitalizations in the intensive care unit: Machine learning approach," in *Proc. 42nd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2020, pp. 5442–5445.

[36] W. McKinney, *Python for Data Analysis: Data Wrangling With Pandas, NumPy, and IPython*. Sebastopol, CA, USA: O'Reilly Media, 2012.

[37] K. Potdar, T. S. Pardawala, and C. D. Pai, "A comparative study of categorical variable encoding techniques for neural network classifiers," *Int. J. Comput. Appl.*, vol. 175, no. 4, pp. 7–9, Oct. 2017.

[38] K. Ho and P. Scott, "Zeta: A global method for discretization of cotitinuous variables," in *Proc. 3rd Int. Conf. Knowl. Discovery Data Mining*, 1997, pp. 191–194.

[39] E. J. Clarke and B. A. Barton, "Entropy and MDL discretization of continuous variables for Bayesian belief networks," *Int. J. Intell. Syst.*, vol. 15, no. 1, pp. 61–92, Jan. 2000.

[40] A. Gupta, T. Liu, and S. Shepherd, "Clinical decision support system to assess the risk of sepsis using tree augmented Bayesian networks and electronic medical record data," *Health Informat. J.*, vol. 26, no. 2, pp. 841–861, Jun. 2020.

[41] A. Tabaie, F. H. Chokshi, A. L. Holder, and S. N. Nemati, "Doubly-robust estimation of effect of imaging resource utilization on discharge decisions in emergency departments," in *Proc. 40th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2018, pp. 3256–3259.

[42] Y.-S. Chen, C.-H. Cheng, C.-J. Lai, C.-Y. Hsu, and H.-J. Syu, "Identifying patients in target customer segments using a two-stage clustering-classification approach: A hospital-based assessment," *Comput. Biol. Med.*, vol. 42, no. 2, pp. 213–221, Feb. 2012.

[43] B. Alsinglawi, O. Alshari, M. Alorjani, O. Mubin, F. Alnajjar, M. Novoa, and O. Darwish, "An explainable machine learning framework for lung cancer hospital length of stay prediction," *Sci. Rep.*, vol. 12, no. 1, pp. 1–10, Jan. 2022.

[44] T. Zebin, S. Rezvy, and T. J. Chaussalet, "A deep learning approach for length of stay prediction in clinical settings from medical records," in *Proc. IEEE Conf. Comput. Intell. Bioinf. Comput. Biol. (CIBCB)*, Jul. 2019, pp. 1–5.

[45] J. P. Allard, H. Keller, K. N. Jeejeebhoy, M. Laporte, D. R. Duerksen, L. Gramlich, H. Payette, P. Bernier, B. Davidson, A. Teterina, and W. Lou, "Decline in nutritional status is associated with prolonged length of stay in hospitalized patients admitted for 7 days or more: A prospective cohort study," *Clin. Nutrition*, vol. 35, no. 1, pp. 144–152, Feb. 2016.

[46] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 785–794.

[47] R. E. Wright, "Logistic regression," in *Reading and Understanding Multivariate Statistics*. Washington, DC, USA: American Psychological Association, 1995, pp. 217–244. [Online]. Available: https://psycnet.apa.org/record/1995-97110-007

[48] G. K. Jha. (2007). *Artificial Neural Networks and Its Applications*. IARI, New Delhi, India. [Online]. Available: https://girishiasri@rediffmail.com

[49] S. E. Gerard, T. J. Patton, G. E. Christensen, J. E. Bayouth, and J. M. Reinhardt, "FissureNet: A deep learning approach for pulmonary fissure detection in CT images," *IEEE Trans. Med. Imag.*, vol. 38, no. 1, pp. 156–166, Jan. 2019.

[50] J. Brownlee. *How to Report Classifier Performance With Confidence Intervals*. Accessed: 2021. [Online]. Available: https://machinelearningmastery.com/report-classifier-performance-confidence-intervals/

[51] A. Rezaei, R. Fathony, O. Memarrast, and B. Ziebart, "Fairness for robust log loss classification," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 4, pp. 5511–5518.

[52] *ExplainerDashboard*. Accessed: 2021. [Online]. Available: https://explainerdashboard.readthedocs.io/en/latest/

[53] *TreeExplainer*. Accessed: 2021. [Online]. Available: https://shap-lrjball.readthedocs.io/en/latest/generated/shap.TreeExplainer.html

[54] E. Zihni, V. I. Madai, M. Livne, I. Galinovic, A. A. Khalil, J. B. Fiebach, and D. Frey, "Opening the black box of artificial intelligence for clinical decision support: A study predicting stroke outcome," *PLoS ONE*, vol. 15, no. 4, Apr. 2020, Art. no. e0231166.

[55] University of Notre Dame. *Lift Charts*. Accessed: 2021. [Online]. Available: https://www3.nd.edu/~busiforc/handouts/DataMining/Lift%20Charts.html

[56] K.-H. Yu and I. S. Kohane, "Framing the challenges of artificial intelligence in medicine," *BMJ Quality Saf.*, vol. 28, no. 3, pp. 238–241, Mar. 2019.

[57] E. H. Shortliffe and M. J. Sepúlveda, "Clinical decision support in the era of artificial intelligence," *J. Amer. Med. Assoc.*, vol. 320, no. 21, p. 2199, Dec. 2018.

[58] B. Alsinglawi, "Predictive analytics framework for electronic health records with machine learning advancements: Optimizing hospital resources utilization with predictive and epidemiological models," Ph.D. thesis, Western Sydney Univ., Penrith, NSW, Australia, 2022. [Online]. Available: https://researchdirect.westernsydney.edu.au/islandora/object/uws:67523/

BELAL S. ALSINGLAWI received the Master of Science and Ph.D. degrees in computer science from Western Sydney University, in 2017 and 2022, respectively. He is currently a Senior Research Fellow at the Swinburne University of Technology. Previously, he worked on the Internet of Things (IoT)-based machine learning project as a Postdoctoral Fellow with The University of Sydney. He has worked and collaborated with key industrial stakeholders on multiple research projects and industrial fellowships with Australian universities, focusing on the IoT, artificial intelligence, and cybersecurity. His work involves applying cutting-edge techniques and transferring research outcomes to solve real-world problems. His extensive professional experience in IT also includes roles as an IT Systems Engineer and a Data Consultant of various organizations, and involvement in digital health and data analytics projects.

FADY ALNAJJAR received the degrees in computer engineering, artificial intelligence, and intelligent systems design engineering from the institutions in United Arab Emirates (UAE) and Japan. As an Associate Professor with UAEU, he has established the AI and Robotics Laboratory and developed AI curricula in the college of IT. With a passion for studying human behavior, he delves into the intricacies of the brain's neural dynamics and cognitive functions; and neuromuscular strategies in learning, adaptation, and recovery. His expertise in this field enables him to pioneer advancements in bio-inspired AI technologies, which are crucial for developing adaptable robotic assistive devices and a range of autonomous assessment systems.

MOHAMMED S. ALORJANI received the M.B.B.S. degree in medicine from the Faculty of Medicine (FoM), Jordan University of Science and Technology (JUST), Irbid, Jordan, in 2002. During the undergraduate study, he received an Academic Distinction from the Medical School, in 1996/1997, and his name was included in the honorary list of JUST. He received a higher specialization degree certificate from JUST, in June 2008, after completing four years of academic residency training (higher specialty training) in pathology with the King Abdullah University Hospital (KAUH)/JUST, Irbid. In 2010, he was granted a sponsorship from JUST to complete fellowship (sub specialization) training in pathology (field: musculoskeletal pathology–bone and soft tissue pathology) with the Royal National Orthopaedic Hospital NHS Trust/UCL, London, U.K., from 2010 to 2012. He has participated and attended many scientific conferences nationally and worldwide, with active presentations/poster participation in some of them as per CV. He has many research contributions during his residency training, fellowship training, and his work as an Academic Staff Member with JUST. He has published many scientific articles as a first author or co-author in high caliber journals. His research fields of interests include pathology, oncology, molecular pathology, epidemiology, artificial intelligence, and machine learning.

OSAMA MOHAMMED AL-SHARI completed a medical training with the Khyber Medical College, Pakistan. Further honed his expertise with a Jordanian Board Certification in Medical Oncology, in 2016, alongside multiple European qualifications, including an Interuniversity Belgian Diploma and a European Diploma in Medical Oncology. He is a Medical Oncologist and an Internist, currently a Consultant Medical Oncologist. He has extensive experience in bone marrow transplantation and hematology, having trained with prestigious institutions, such as UZ Brussels, and the Catholic University of Leuven, Belgium. He is actively involved in teaching medical students with Jordan University of Science and Technology and has participated in numerous international oncology conferences. His clinical research includes significant contributions to international clinical trials.

MAURICIO NOVOA MUNOZ is currently a Lecturer in industrial design with the School of Engineering, Design, and Built Environment, Western Sydney University, Australia. He is a Senior Fellow of the Higher Education Academy, U.K., and the Academic Program Advisor for industrial design with Western Sydney University. With over three decades in the industry and 18 years in academia, he focuses on social transformation through innovation in design. He supervises Ph.D. candidates across several interdisciplinary fields related to industrial design and technology. His research interests include design methodologies, digital transformation, and technology diffusion. He is an Active Member of international design committees and a frequent speaker at global events.

OMAR MUBIN is currently an Associate Professor, a Senior Academic, and a Researcher in human–computer interaction with the School of Computer, Data and Mathematical Sciences, Western Sydney University, Australia. He is also the Associate Dean of high degree research with his faculty. He is involved in teaching and supervising (undergraduate and postgraduate) students in the broader area of human–computer interaction, mobile computing, and health informatics. His primary research interests include human–robot interaction, human-agent interaction, and scientometric. Specifically, he studies social robotics and their applications and consequently interaction with humans in education, public spaces, and information dissemination scenarios.