POLYPUBLIE
Polytechnique Montréal

POLYTECHNIQUE MONTRÉAL
UNIVERSITÉ D'INGÉNIERIE

Titre: Title:	A comprehensive review of ICU readmission prediction models: From statistical methods to deep learning approaches
Auteurs: Authors:	Waleed Sayed Ahmed Fathy Gharib, Guillaume Émériaud, & Farida Cheriet
Date:	2025
Type:	Article de revue / Article
Référence: Citation:	Gharib, W. S. A. F., Émériaud, G., & Cheriet, F. (2025). A comprehensive review of ICU readmission prediction models: From statistical methods to deep learning approaches. Artificial Intelligence in Medicine, 103126 (16 pages). https://doi.org/10.1016/j.artmed.2025.103126

Document en libre accès dans PolyPublie
Open Access document in PolyPublie

URL de PolyPublie: PolyPublie URL:	https://publications.polymtl.ca/64598/
Version:	Version officielle de l'éditeur / Published version Révisé par les pairs / Refereed
Conditions d'utilisation: Terms of Use:	Creative Commons Attribution-Utilisation non commerciale 4.0 International / Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC)

Document publié chez l'éditeur officiel
Document issued by the official publisher

Titre de la revue: Journal Title:	Artificial Intelligence in Medicine
Maison d'édition: Publisher:	Elsevier BV
URL officiel: Official URL:	https://doi.org/10.1016/j.artmed.2025.103126
Mention légale: Legal notice:	Under a Creative Commons license (http://creativecommons.org/licenses/by-nc/4.0/)

Ce fichier a été téléchargé à partir de PolyPublie, le dépôt institutionnel de Polytechnique Montréal
This file has been downloaded from PolyPublie, the institutional repository of Polytechnique Montréal

https://publications.polymtl.ca

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

journal homepage: www.elsevier.com/locate/artmed

A comprehensive review of ICU readmission prediction models: From statistical methods to deep learning approaches

Waleed Fathy [a,c,*], Guillaume Emeriaud [b], Farida Cheriet [a]

[a] Department of Computer and Software Engineering, Polytechnique Montréal, Montreal, Quebec, Canada
[b] Department of Pediatrics, CHU Sainte-Justine, Université de Montréal, Montreal, Quebec, Canada
[c] Department of Electronic and Communication Engineering, Zagazig Univeristy, Zagazig, Sharkia, Egypt

ARTICLE INFO

Keywords:
Intensive care unit
Readmission
Review
Machine learning
Deep learning

ABSTRACT

The prediction of Intensive Care Unit (ICU) readmission has become a crucial area of research due to the increasing demand for ICU resources and the need to provide timely interventions to critically ill patients. In recent years, several studies have explored the use of statistical, machine learning (ML), and deep learning (DL) models to predict ICU readmission. This review paper presents an extensive overview of these studies and discusses the challenges associated with ICU readmission prediction. We categorize the studies based on the type of model used and evaluate their strengths and limitations. We also discuss the performance metrics used to evaluate the models and their potential clinical applications. In addition, this review explores current methodologies, data usage, and recent advances in interpretability and explainable AI for medical applications, offering insights to guide future research and development in this field. Finally, we identify gaps in the current literature and provide recommendations for future research. Recent advances like ML and DL have moderately improved the prediction of the risk of ICU readmission. However, more progress is needed to reach the precision required to build computerized decision support tools.

Contents

https://doi.org/10.1016/j.artmed.2025.103126
Received 13 June 2023; Received in revised form 4 October 2024; Accepted 29 March 2025
Available online 16 April 2025
0933-3657/© 2025 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).

Table 1
Summary of PICO elements for review paper.

PICO element	Inclusion criteria	Exclusion criteria
P (population)	- Adult patient - Discharged alive from ICU	- Non-ICU patients - Died in the ICU - Age < 15
I (Intervention)	Studies involving the development or evaluation of predictive models for ICU readmission	- Diagnosis or cohort specific - Exploring risk factors only - Exploring readmission rate only - Medical studies without predictive models
C (comparison)	Any	No restriction
O (outcome)	Studies that predict ICU readmission within 24 h or more during the same hospitalization	- Not ICU readmission - ICU readmission after hospital discharge

1. Introduction

Intensive Care Unit (ICU) readmission is defined as a patient's nonscheduled return to ICU within a short prespecified period after discharge. This is a crucial risk factor, which ranges between 4% and 14%, associated with increased morbidity and mortality rates, longer hospital stays, and higher healthcare costs. A two to ten-fold higher mortality rate is observed, compared to patients who do not experience ICU readmission. Preventing unplanned readmissions to the ICU has been identified as a key objective in quality improvement strategies to minimize avoidable risks [1].

Electronic Health Record (EHR) data offer a vast resource of information on ICU patients, making it suitable for a data-driven approach to tackle unplanned readmissions. The complex analysis of the patient data collected in the ICU makes clinical prediction challenging. A huge amount of information is available, and it is hard for the human brain to analyze all of them at the same time. In addition, EHR data have challenges like high missing rates and imbalanced outcome classes. Machine learning (ML) algorithms, including deep learning (DL), have shown great success in medical diagnosis and decision-making, utilizing numerous features that human analysis cannot handle [2].

Only 7 review papers have looked at ICU readmission [3–9], but these systematic reviews only include three to five studies, which does not represent the full scope of research on ICU readmission risk prediction models. Even in [9], which reviewed 33 studies, the focus was primarily on assessing the risk of bias in prediction models rather than on model development; moreover, it included studies targeting ICU readmission within specific cohorts or diagnoses. Therefore, further studies are needed to provide a more comprehensive understanding of the topic that can help researchers develop more effective models and avoid previous shortcomings. This work provides an extensive review of research papers that have utilized predictive models (statistical methods, ML, and DL) to develop models for predicting ICU readmission. Following established medical Artificial Intelligence (AI) life cycle guidelines [10], encompassing model development, data creation, and AI safety, our systematic literature review aims to:

- Summarize previous research to discuss the models used to handle different types of data, such as demographic data, temporal data, medical codes, and clinical notes.
- Systematically evaluate and benchmark the performance of the models to ensure their effectiveness in real-world applications.
- Assess the interpretability of the models.
- Explore various preprocessing techniques employed in data analysis, encompassing adept imputation methods for managing missing data, strategies to rebalance databases, and meticulous feature selection methodologies designed to mitigate redundancy.
- Show the databases utilized, the size of the study cohort, and the rate of readmissions.
- Examining challenges and perspectives in achieving seamless integration of AI-driven approaches for predicting ICU readmission.

The remaining sections of the paper are structured as follows: Section 2 will discuss the criteria used to select relevant research papers, Section 3 will analyze and discuss the selected research papers, Section 4 will discuss the current limitations and challenges faced by researchers in this area, Section 5 will present potential future directions to advance the field, and Section 6 will provide a conclusion to this review.

2. Methods

We followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) framework to report our findings [11]:

2.1. Eligibility criteria

In this literature review, we included the peer-reviewed papers that focused on predicting ICU readmission during the same hospitalization using a predictive model (statistical methods, ML models, or DL models). The selection and screening of papers were conducted collaboratively by the first two authors and the eligibility were conducted with the first author only. If any disagreements had arisen, we would have sought a formal discussion or a third-party review by the third author to reach consensus. Table 1 shows the Population, Intervention, Comparison, and Outcome (PICO) framework for this work.

2.2. Data source and searching strategy

Four search engines, namely Google Scholar, PubMed, Compendex, and Wiley Online Library, were utilized to conduct the literature review. Only English-language research papers that were published in reputable peer-reviewed conferences or journals up to December 2023 were considered. The following search query was used to find the required publications: (((((Readmission OR Re-admission OR readmi*) AND(Artificial* OR "Machine learning" OR "Deep learning" OR "Neural network" OR "Predict* model* ") AND (ICU OR "Intensive care Unit")))) AND (English WN LA)), and it was adapted for each search engine.

* Corresponding author at: Department of Computer and Software Engineering, Polytechnique Montréal, Montreal, Quebec, Canada.
E-mail addresses: wfathy.gharib@polymtl.ca (W. Fathy), guillaume.emeriaud.med@ssss.gouv.qc.ca (G. Emeriaud), farida.cheriet@polymtl.ca (F. Cheriet).

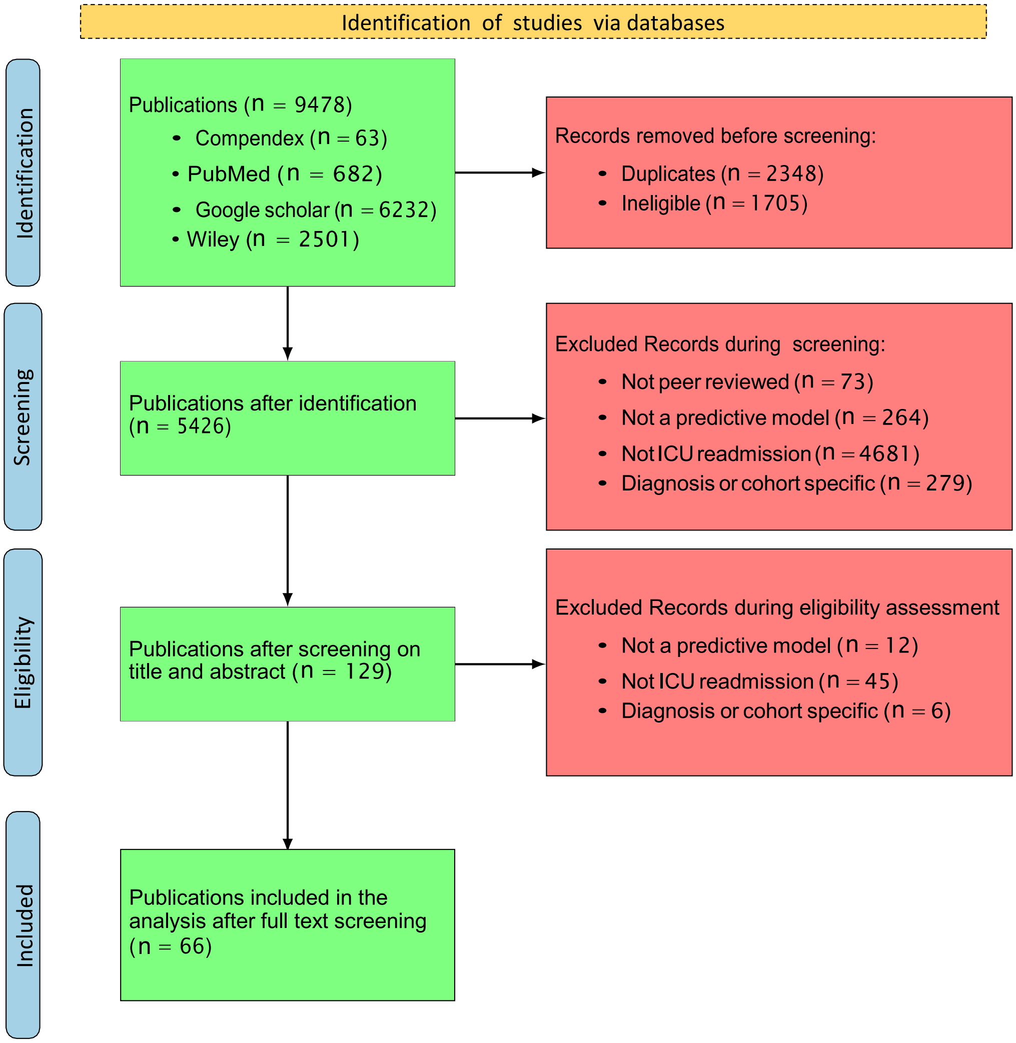
Identification of studies via databases
Identification
Screening
Eligibility
Included
Publications (n = 9478)
• Compendex (n = 63)
• PubMed (n = 682)
• Google scholar (n = 6232)
• Wiley (n = 2501)
Records removed before screening:
• Duplicates (n = 2348)
• Ineligible (n = 1705)
Publications after identification (n = 5426)
Excluded Records during screening:
• Not peer reviewed (n = 73)
• Not a predictive model (n = 264)
• Not ICU readmission (n = 4681)
• Diagnosis or cohort specific (n = 279)
Publications after screening on title and abstract (n = 129)
Excluded Records during eligibility assessment
• Not a predictive model (n = 12)
• Not ICU readmission (n = 45)
• Diagnosis or cohort specific (n = 6)
Publications included in the analysis after full text screening (n = 66)

Fig. 1. PRISMA Workflow: A Visual Representation of the Systematic Review Process.

2.3. Study selection

After retrieving 9478 papers and filtering out 4053 duplicates or ineligible studies, 5426 papers were screened based on criteria such as ICU readmission as an outcome and the use of predictive models. Ultimately, 66 papers met our inclusion criteria and were included in the final review. The selection process for including research papers was illustrated using a PRISMA Workflow Fig. 1.

2.4. Data collection and analysis

We performed a quantitative descriptive analysis of the papers using CHecklist for critical Appraisal and data extraction for systematic Reviews of prediction Modelling Studies (CHARMS) [12]. Our analysis covered ICU readmission period, database, data type, preprocessing, classifier, interpretability, and performance, with the extracted data summarized in tables.

3. Results

3.1. Feature representation techniques

Effective feature representation is paramount in predicting ICU readmission due to the rich information in EHR data. Most statistical and ML models for predicting ICU readmission used primarily demographic data and temporal data, with some studies also exploring the use of medical codes and clinical notes. To use temporal data, statistical features were extracted using techniques such as descriptive statistics and entropy [13–22]. These features fail to capture detailed temporal trends. Techniques like frequent subgraph mining were used to capture temporal trends [23,24] but it overlooks critical patterns in understanding physiological measurement dynamics over time. To overcome this, some studies converted temporal data into qualitative variables [25–30]. However, these methods simplify complex temporal patterns into discrete categories, neglecting time embedding in a vector space. Medical codes, especially diagnoses, were simplified in several studies using scores such as APACHE [31] and SOFA [32] scores [19,26,33,34] but lacked the rich context and specificity required for accurate prediction. Other studies focused on using pre-trained embeddings based on International Classification of Diseases (ICD) codes such as Choi [35] and the Clinical Classification Software (CCS) [36] or clustering similar diagnoses [17,21,22,25,37–39]. Treatments were represented by several features such as usage duration [18,20,40,41]. Text data were represented in several ways based on Bag-of-Words (BoW) and sting matching [42–44]. Despite some effectiveness, these methods struggled with the complexity of textual data.

These techniques do not capture the potential of these data. DL models offer automated feature learning, integrating these data to capture complex patterns and dependencies. Learning representations using Long Short-Term Memory (LSTM), convolutional Neural Network (CNN), and Neural Ordinary Differential Equations (ODEs) was used in several studies [21,30,38,45]. However, these representations face challenges due to sparse data, different modalities, variable lengths of hospital stay, and irregular time intervals. To address this, others proposed embedding time-related information to code embeddings using time-aware attention or exponential time-decay functions [30,46].

Natural Language Processing (NLP) techniques are crucial for enhancing information extracted from clinical notes. Word embedding techniques such as Word2Vec [47] and BioWordVec [48] were used to capture semantic meanings in a continuous space, grouping similar words closely [49]. With the introduction of transformer-based large language models (LLMs), especially Bidirectional Encoder Representations from Transformers (BERT) [50], the representation of textual data has undergone significant improvement as it can capture bidirectional contextual information. BioBERT [51] and ClinicalBERT [52] are both specialized versions of BERT, designed for processing biomedical and clinical text, respectively [29,39]. However, using these word embedding techniques is ineffective due to the lengthy and noisy content of clinical notes. In addition, they overlook the graphical structure mirroring the physician's decision-making process and lack interpretability, potentially hindering performance.

Recent approaches focus on effectively extracting entities from clinical notes and medical codes, and embedding them with meaningful medical representations. One approach embedded ICD-9 codes into a hyperbolic space [53] using Poincaré embeddings to represent hierarchical structures [54]. Other approaches integrated external knowledge from medical ontologies or domain-specific databases such as the Unified Medical Language System (UMLS) [55] and the Systematized Nomenclature of Medicine-Clinical Terms (SNOMED CT) [56], or created a graph representation of patient encounters, or a combination of both approaches [27,57,58]. Enriching EHR data with external knowledge and generating graph embeddings improves DL performance and helps ML models better utilize clinical data.

Graphs are a flexible framework for incorporating data from multiple modalities. Most graph-based models have focused on representing either medical codes [59–63] or clinical notes [64]. However, exploring optimal fusion strategies for different modalities has emerged as a promising research direction in patient representation learning [65–68]. Variational regularization was proposed to handle irregular time intervals between patient encounters [62]. Graphs are limited by fuzzy patient data relevance and struggle for higher performance due to noise from diverse disease types. Leveraging external medical knowledge to integrate domain-specific information enhances node representation and facilitates a better understanding of the relationships among different entities [61,64,66–68]. Using contrastive learning techniques [69] to refine graph representations by generating pairs of nodes to ensure consistency, effectively bringing embeddings of patients with the same label closer in the embedding space [63,67]. These methods enhance the ability to model intricate relationships within EHR data, leading to improved performance and clinical insights.

3.2. Model types

This section reviews statistical, machine learning, and deep learning models used to predict ICU readmission.

3.2.1. Statistical approaches

Table 2 identifies 21 studies, that used various statistical approaches, including Logistic Regression (LR), fuzzy clustering, Dynamic Linear Models (DLMs), and Conditional Random Fields(CRF), have been employed to develop ICU readmission prediction models.

Logistic Regression (LR) models the probability of a binary outcome by applying a logistic function to the linear combination of predictor variables. Campbell, et al. [33] initially achieved moderate Area Under the Curve (AUC) of 62% and 67% for predicting readmissions using LR. In [70], they proposed a scoring tool by normalizing LR coefficients achieving an AUC of 76%. Subsequent improvements included feature selection techniques, with Jo, et al. [71] achieving an AUC of 76% through the Likelihood Ratio-Test (LR-Test), and Frost, et al. [34] reaching 66% with backward-deletion. To further enhance performance, different feature extraction techniques were employed. Badwi, et al. [37] grouped diagnoses into 26 categories, extracted statistical features, used Multiple Imputation by Chained Equations (MICE) [72] for missing data, and selected significant features with LR-test and stepwise LR, achieving an AUC of 71%. Ouanes, et al. [26] transformed temporal variables into qualitative ones and identified significant predictors through multivariate analysis, resulting in an improved AUC of 74%. Xue, et al. [23] achieved an AUC of 66% using adapted the Subgraph Augmented Non-negative Matrix Factorization (SANMF) but with limited impact on overall trend correlation. Utilizing clinical notes, in [57], they reached an AUC of 75% using UMLS to identify medical concepts and assigning unique Concept Unique Identifiers (CUIs), generating a Bag-of-CUIs, while in [73], they attained 76% with BOW embedding. The performance of LR-based models was moderate, with AUC values ranging from 62% to 76%, and an average AUC of 70%.

Fuzzy clustering found application in multiple studies. It allows data points to belong to multiple clusters simultaneously. The Takagi–Sugeno fuzzy model (TSFM) is a fuzzy inference system that uses linear functions associated with fuzzy sets to represent rules, activated based on input values, producing a weighted average of outputs to model complex systems efficiently [74]. Techniques such as fuzzy c-means (FCM) [75], mixed fuzzy clustering (MFC) [76], and probabilistic fuzzy systems (PFS) [77] were used to determine the antecedent fuzzy sets and the number of rules of TSFM. FCM optimizes the sum of squared differences for numerical data, MFC enhances FCM for spatiotemporal data, while PFS merges fuzzy logic with probabilistic reasoning, integrating linguistic system descriptions with statistical data properties.

Fialho et al. [13] achieved an AUC of 72% using a TSFM model based on FCM clustering and six variables, and an AUC of 66% with a TSFM model based on PFS with FCM and the nearest neighbor heuristic for membership functions [14]. Vieira, et al. [78] suggested improving this by merging numerical and medical text annotations, and Curto, et al. [43] further advanced this with a Fuzzy FingerPrint (FFP) classifier [79] and Pareto-inspired membership function [80], achieving an AUC of 80% with text data. In [15], temporal data were represented using Shannon entropy and weighted average, with feature selection achieved through Binary Fish School Search (BFSS) [81] combined with FCM, resulting in an AUC of 69%. In [82], three TSFM model approaches were used: FCM, modified MFC for fixed-length multivariate time series, and MFC-FCM with feature transformation, with MFC and FCM outperforming MFC-FCM and achieving an AUC of 58%. A follow-up [83] extended MFC to handle unequal-length multivariate time series, using FCM for transformation and MFC grouping (FCM-U^{MFC}), achieving an AUC of 64%.

Fernandes, et al. [84] suggested using ensemble learning with FCM-based clustering to generate patient clusters for TSFM models development. They employed classifier selection techniques including a priori decisions based on cluster center distances to patient characteristics and a posteriori decisions to select outcomes with lower uncertainty. The posteriori decision outperformed the a priori approach, resulting in an AUC of 75%. A follow-up [85] found no significant difference between aggregation techniques and classifier selection. Viegas et al. [16] enhanced this approach by using Sequential Forward Selection (SFS) for feature selection and Gustafson-Kessel (GK) clustering [86], combining sensitivity and specificity models with weights based on uncertainty, achieving an AUC of 77%. Overall, fuzzy-based models performed similarly to LR-based models, with AUCs ranging from 58% to 81% and an average of 71%.

DLMs are valuable for modeling time series data, incorporating both system dynamics and observation uncertainty. They model time series

Table 2
Statistical approaches for ICU prediction (same*: same hospitalization, ** 🟢: used and 🔴: unused).

Study	Database name (Sample size)	Time period (days) Readmission rate %	Data type				Classifier	Performance	
			Demographic	Temporal	Clinical notes	Medical codes		AUC (%)	Accuracy (%)
Campbell 2008 [33]	Private (6208)	same* (9) 2 (3)	🟢**	🔴	🔴	🔴	LR	62 67	–
Haribhakti 2021 [70]	Private (883)	same (9)	🟢	🔴	🔴	🔴	LR	76	–
Jo 2015 [71]	Private (343)	same (10)	🟢	🟢	🔴	🔴	LR	76	–
Forst 2010 [34]	Private (14,952)	same (7)	🟢	🔴	🔴	🔴	LR	66	–
Badwi 2012 [37]	eICU (704,963)	2 (3)	🟢	🟢	🔴	🔴	LR	71	–
Ouanes 2012 [26]	Private (3462)	7 (2)	🟢	🔴	🔴	🔴	LR	74	–
Xue 2019 [23]	MIMIC-II (1170)	30 (27)	🟢	🟢	🔴	🟢	LR	66	–
Li 2019 [57]	MIMIC-III (45,305)	30 (5)	🔴	🔴	🟢	🔴	LR	75	–
Moerschbacher 2023 [73]	MIMIC-III (4522)	30 (50)	🔴 🟢	🔴 🟢	🟢 🟢	🔴 🟢	LR RF	76 70	69 71
Fialho 2012 [13]	MIMIC-II (1028)	3 (13)	🔴	🟢	🔴	🔴	FCM	72	71
Fialho 2013 [14]	Private (3271)	3	🟢	🟢	🔴	🔴	PFS	66	67
Vieira 2013 [78]	MIMIC-II (1028)	3 (13)	🔴 🔴	🟢 🟢	🔴 🟢	🔴 🔴	TSFM FFP	72 –	– –
Curto 2016 [43]	MIMIC-II (12,091)	3 (6)	🔴 🔴	🔴 🟢	🟢 🔴	🔴 🔴	FFP TSFM	80 64	84 69
Sargo 2014 [15]	MIMIC-II (726)	3 (12)	🔴	🟢	🔴	🔴	FCM	69	56
Ferreira 2015 [82]	MIMIC-II (2653)	3 (8)	🟢	🟢	🔴	🔴	MFC	58	59
Salgado 2016 [83]	MIMIC-II (1389)	3 (10)	🟢	🟢	🔴	🔴	FCM-U^{MFC}	64	57
Fernandes 2014 [84]	MIMIC-II (1010)	3 (13)	🔴	🟢	🔴	🔴	FCM posteriori	75	70
Salgado 2015 [85]	MIMIC-II (1010)	3 (13)	🔴	🟢	🔴	🔴	FCM posteriori	81	72
Viegas 2017 [16]	MIMIC-II (1499)	3 (7)	🟢	🟢	🔴	🔴	GK	77	71
Caballero 2015 [27]	MIMIC-II (11,648)	30	🔴	🟢	🟢	🔴	DLMs	93	–
Venugopalan 2017 [28]	MIMIC-II (32,331)	30 (24)	🟢	🟢	🔴	🔴	CRF	MCC: 73	90

by estimating the state vector from observed data, using the Kalman filter to update the estimate with new observations [87]. Caballero, et al. [27] used DLMs with Bayesian time series to capture temporal dependencies and update readmission predictions, extracting features focusing on clinically named entities using UMLS and statistical topic modeling, achieving an AUC of 93%.

CRFs are probabilistic models used for labeling sequential data. They model the conditional probability of labels given observations, capturing complex dependencies [88]. Venugopalan, et al. [28] used k-means and FCM for imputing temporal data, employed CRF for capturing time-varying dynamics, and applied LR and Neural Network (NN) for static data classification. CRF with FCM-based imputation achieved 90% accuracy, but models combining with k-means imputation surpassed CRF alone (80%), reaching 87% accuracy.

Table 2 summarizes statistical approaches, revealing AUC scores ranging widely from 58% to 93%. Despite high AUCs, the models' reliability is limited due to the small patient sample size.

3.2.2. Machine learning approaches

Table 3 highlights 22 studies using various machine learning models, primarily supervised approaches like Decision Trees (DT), with some exploring unsupervised methods like K-means Clustering.

NN learns complex patterns through interconnected nodes using feedforward flow, activation functions, and back-propagation [89]. In [90], NN achieved AUCs of 87% and 79% on complete and subsampled datasets, respectively, showing that the dataset's histogram had an improper distribution for probability models and no specific feature selection method was recommended. Junqueira, et al. [91] used Symmetrical Uncertainty (SU) [92] for feature selection, finding NN performed best with a 64% AUC on a temporally split, sub-sampled dataset, showing consistent risk factors over time. In [93], four ML models performed well, with the NN model achieving the best overall performance with an F1-score of 84%. The performance of NN-based models was moderate.

Support Vector Machine (SVM) uses a hyperplane to separate different class data points. It excels in high-dimensional spaces and handles non-linear data through the kernel trick [94]. Negar, et al. [44] used BOW to create a document-term matrix from clinical notes, achieving an AUC of 71% and 74% with feature selection using SVM, outperforming other ML models on a sub-sampled database. Naive Bayes (NB) is a simple probabilistic classifier based on applying Bayes' theorem with strong (naive) independence assumptions between the features [95]. In [96], NB outperformed other ML models on an oversampled database, achieving an accuracy of 99%. Oversampling by duplicating minority class data without proper separation between training and test sets leads to biased metrics and inflated specificity scores, with the model showing a significantly lower specificity of 72% without oversampling.

DT is a tree-like model that uses a flowchart-like structure of decisions and their possible consequences to make predictions or decisions [97]. Random Forest (RF) improves DT by addressing overfitting through ensemble learning. It combines predictions from multiple trees, offering enhanced robustness, and feature importance [98]. DT is interpretable, while RF excels with larger datasets. In [99], a DT algorithm based on axiomatic fuzzy set theory (AFS-DT) [100] achieved an AUC of 61% using Cohen's kappa coefficient as a fitness function, with similar performance (60% AUC) when trained on only six variables recommended in [13], indicating no significant loss of information. In [17], a proposed noise reduction learning (NRL) system tackled data sparsity by under-sampling overlapping points with k-nearest neighbors (KNN). Ensemble learning using DT and LR models with L1 regularization achieved an AUC of 81%. In [42], a weight decay RF model was utilized for both imputation and dataset rebalancing, leveraging features from temporal data and clinical notes to achieve an AUC of 88%, surpassing other machine learning models. Alghatani, et al. [101] used ICU data from the first day and applied the percent point function to determine quantile thresholds. SVM achieved an AUC of 59% across the entire dataset, while RF reached a best AUC of 74% for the sub-sampled dataset.

In [102], the patient forest model was introduced, leveraging an ensemble of DTs within the gcForest framework. This approach utilized multi-grained scanning and cascade forest modules to process EHR encounters and extract features at various granularities, which were then refined through multiple levels of RFs. The model, trained with optimized convolutional filter and RF parameters via backpropagation to minimize binary cross-entropy loss, achieved an AUC of 87%. Its superior performance was attributed to its ability to capture the heterogeneity and complexity of patient data through patient-specific DTs. The aggregation of predictions from multiple trees also reduced variance and enhanced result stability compared to models using a single global classifier.

Boosting techniques like AdaBoost, Gradient Boosting (GBM), XGBoost, and LightGBM enhance predictive performance by emphasizing misclassified instances sequentially. AdaBoost is effective and less prone to overfitting but sensitive to noise [103]. GBM ensures high accuracy but is computationally intensive [104]. XGBoost improves Gradient Boosting with regularization and parallel processing, enhancing computational efficiency [105]. LightGBM employs histogram-based learning for efficient processing of large datasets, although it may necessitate careful parameter tuning [106]. In [24], a GBM model trained on sub-sampled data from the University of Chicago and MIMIC-III databases achieved AUCs of 76% and 71%, respectively. In [18], GBM outperformed several ML models with an AUC of 85% using the STARR database but faced a decline to 60% during external validation on the MIMIC-IV dataset, suggesting that GBM may be less prone to overfitting than LR. Zhu, et al. [40] applied transfer learning from MIMIC-II to CHOA databases. They applied non-negative matrix factorization (NMF) and convolutional autoencoders (CAE) for feature representation, and fine-tuned classifiers on CHOA data. Although transfer learning improved GBM performance with CAE-extracted features to an AUC of 62%, a GBM classifier achieved an AUC of 77% when trained directly on CHOA data.

In [19], AdaBoost achieved the best performance with an AUC of 91% using the arrival attribute set. Cost-sensitive classification outperformed Synthetic Minority Over-sampling Technique (SMOTE) [107] for class imbalance, and the study highlighted the value of early patient characteristics over data available at ICU discharge. The study's limited private ICU sample reduces variability and generalizability, and manual data entry introduces error risk, while modeling time-varying variables as categorical data neglects their dynamic nature. Desautels, et al. [108] developed AutoTriage ML, which utilized transfer learning with AdaBoost, trained on MIMIC-III (source) and CUH (target) datasets. By adjusting the weighting between datasets, this approach achieved an AUC of 71%, demonstrating that prioritizing the target dataset in transfer learning outperformed models trained solely on the source or target datasets served as an effective regularizer, preventing overfitting to the target domain and enhancing performance. In [25], the LightGBM classifier, trained with features selected for mutual information and weighted for the minority class, achieved an AUC of 79%. The study underscored that diagnoses are key contributors to ICU readmission.

Pakbin, et al. [22] used XGBoost to predict readmissions at different time windows, combining multiple feature entries to reduce missingness. The model outperformed LR, achieving an AUC of 76% and 84% for predicting 72 h and bounceback readmissions, respectively. Their findings highlighted distinctions in short- and long-term readmission risks, with the diagnosis as the most correlated risk factor. Thoral, et al. [20] used SHapley Additive exPlanations (SHAP) [109] to highlight feature importance and enhance the interpretability of an XGBoost model trained on AmsterdamUMCdb data. Feature selection with LR and L1 regularization achieved an AUC of 77%. In a subsequent study [110], Thoral's model was assessed through a temporal validation design Leiden UMC data. External validation revealed moderate discrimination with an AUC of 72%. Retraining the model using different time point subsets improved the AUC to 79%, indicating no changes in data drift affecting performance over time, and highlighting the importance of retraining models on new data. In [111], XGBoost's hyperparameters were optimized using Tree-structured Parzen Estimator (TPE) [112], a Bayesian optimization technique. They achieved remarkable results with an AUC of 92% and an Area Under Precision-Recall Curve (AUPRC) of 65%, with SHAP values identifying length of stay as the most influential feature. A low AUPRC indicates that the model struggles to correctly identify true positives while minimizing false positives, which is particularly problematic in imbalanced datasets where the positive class is rare.

Hegselmann, et al. [41] developed an Explainable Boosting Machine (EBM), a generalized additive model using shape functions for interactions between variables and the logit link for dichotomous classifications. With features selected via mean absolute log-odds score from ANIT-UKM hospital data, the EBM achieved an AUC of 68%, outperforming LR and Recurrent neural Network (RNN) but similar to GBM. Validation on MIMIC-IV, benefiting from better data quality, improved the AUC to 76%. The EBM's transparency, reviewed by a multidisciplinary team, highlights its advantages over black-box models for healthcare applications. Tree-based models showed promising performance with AUCs ranging from 59% to 92%, averaging 77%.

The K-means algorithm partitions patients into clusters by assigning them based on similarity to the cluster centroid and iteratively refining assignments to minimize within-cluster variance [114]. In [113], K-means outperformed k-medoids [115] and x-means [116], achieving a Davies–Bouldin Index (DBI) of 56.

Table 3 summarizes ML approaches. While ML models have significantly improved results, their performance tends to decline with larger sample sizes. Developing more complex models is essential for creating more reliable predictions.

3.2.3. Deep learning approaches

Table 4 identifies 23 studies used to handle temporal data and clinical notes directly using models like RNN [117] or CNN [118] are often used. LSTM [119] and Gated Recurrent Unit (GRU) [120] are RNN variants that capture long-range dependencies in sequential data by retaining past information to improve future predictions. On the other hand, CNNs learn hierarchical representations of data, which are useful for capturing patterns in sequential data with a spatial component.

The LSTM-CNN model was used in many studies. In [21], it achieved an AUC of 79%, outperforming LSTM, CNN, and several ML models trained in statistical features. Their study underscores the LSTM-CNN model's ability to handle time series data with high volatility and unstable conditions effectively. To enhance this work, Zebin, et al. [38] rebalanced the dataset using SMOTE and categorized ICD-9 codes into 17 classes, achieving an AUC of 82%. In [54], the model was enhanced with advanced ICD-9 embeddings, including Poincaré embeddings, which achieved AUCs of 79% and 78% for clinical notes and billing system ICD-9 codes, respectively, outperforming all graph embedding methods except TransE. Poincaré embeddings with 100 dimensions proved notably efficient, achieving an AUC of 72% and demonstrating their effectiveness in hierarchical data representation. Chen, et al. [121] introduced predictive process monitoring (PPM) using event logs to enhance ICU support. PPM learns from historical complete traces and makes predictions for ongoing, incomplete traces, treating each ICU stay as a process trace and utilizing rich time series information. The LSTM-CNN model improved with longer prefix lengths, achieving an AUC of 64% with a prefix length of 21. Overall,

Table 3
ML approaches for ICU prediction.

Study	Database name (Sample size)	Time period (days) Readmission rate %	Data type				Classifier	Performance		
			Demographic	Temporal	Clinical notes	Medical codes		AUC (%)	APURC (%)	Accuracy (%)
Inan 2018 [90]	MIMIC-III (11,000)	–	🟢	🟢	🔴	🔴	NN	87	95	95
Junqueira 2019 [91]	MIMIC-III (42,307)	30 (11)	🟢	🔴	🔴	🔴	NN	64	–	86
Raza 2023 [93]	MIMIC-III (6500)	30 (50)	🟢	🟢	🔴	🟢	NN	F1 score: 84		
Negar 2022 [44]	MIMIC-III (10,894)	30 (30)	🔴	🔴	🟢	🔴	SVM	74	–	–
Braga 2014 [96]	Private	30 (1)	🟢	🟢	🔴	🔴	NB	–	–	99
Silva 2015 [99]	MIMIC-II (19,075)	same (13)	🔴	🟢	🔴	🔴	AFS-DT	59	–	61
He 2022 [17]	MIMIC-II (1622)	3 (1)	🔴	🟢	🔴	🟢	DT+ LR	81	–	73
Wang 2021 [42]	Private (4697)	2 (13)	🟢	🟢	🟢	🔴	RF	88	–	87
Alghatani 2022 [101]	MIMIC-III (44,626)	same (7)	🟢	🟢	🔴	🔴	RF	74	–	68
Khodadadi 2023 [102]	eICU (41,026)	same (17)	🔴	🔴	🔴	🟢	gcForest	87	60	–
Rojas 2018 [24]	Private (24,885)	same (11)	🟢	🟢	🔴	🟢	GBM	76	–	–
	MIMIC-III (42,303)[a]	same (8)						71	–	–
Shi 2022 [18]	Private (3107)	7 (9)	🟢	🟢	🔴	🟢	GBM	85	41	90
	MIMIC-IV (13,841)[a]	7 (6)						60	8	93
Zhu 2022 [40]	MIMIC-II (32,331)	30 (24)	🟢	🟢	🔴	🟢	GBM	77	–	–
	Private (5739)	30								
Loreto 2020 [19]	Private (9926)	same (7)	🟢	🟢	🔴	🔴	RF Adaboost	91	–	–
Desautels 2017 [108]	Private (2018)	2 (4)	🟢	🟢	🔴	🔴	AdaBoost	71	–	–
	MIMIC-III (44,741)	2 (13)								
Fathy 2023 [25]	MIMIC-III (31,151)	3 (3)	🟢	🟢	🔴	🟢	LightGBM	79	–	–
Pakbin 2018 [22]	MIMIC-III (3637)	3 (4)	🟢	🟢	🔴	🟢	XGBoost	76	–	–
		30 (12)						75	–	–
		same (7)						84	–	–
Thoral 2021 [20]	Private (18 034)	7 (4)	🟢	🟢	🔴	🔴	XGBoost	77	–	12
De Hond 2022 [110]	Private (10,052)	7 (6)	🟢	🟢	🔴	🔴	XGBoost	79	–	–
González-Nóvoa 2023 [111]	MIMIC-III (28,557)	same (8)	🟢	🟢	🔴	🔴	XGBoost	92	65	–
Hegselmann 2022 [41]	Private (15,589) [a]	3 (5)	🟢	🟢	🔴	🟢	EBM	68	12	–
	MIMIC-IV (19,108)	3 (7)						76	22	–
Veloso 2014 [113]	Private (1043)	30 (4)	🟢	🟢	🔴	🔴	K-means	DBI: 56		

[a] External validation.

the LSTM-CNN models performed well, with AUC values ranging from 63% to 82%, and an average AUC of 76%.

The attention mechanism assigns dynamic weights to input elements, helping models understand data well and focus on specific parts for better predictions [122]. To address RNN and CNN limitations, the transformer architecture processes all input tokens simultaneously, capturing long-range dependencies, handling irregularly sampled data, and improving text data analysis for identifying high-risk patients. In [49], the patient's continuous clinical notes were concatenated and embedded using Word2Vec. An augmented CNN with a multi-headed attention mechanism was employed to extract problems, allowing for variable text spans while maintaining interpretability. The model explored various problem representations, including rolled-up ICD-9 codes and Phecodes [123], achieving AUCs of 71% and 69% for bounceback and 30-day readmission, respectively. The model's interpretability is enhanced by using the intermediate problem list for final predictions. In [45], attention mechanisms in LSTM were examined for their correlation with feature importance and alterability. Log-odds attention showed a modest correlation and had minimal impact on predictions, raising questions about the explanatory power of attention mechanisms. Although log-odds attention offered interpretation, it differed from learned attention distributions. Despite this, incorporating attention mechanisms into LSTM consistently improved AUC, reaching 71% with either additive or log-odds attention.

Longformer is a transformer model that handles longer sequences of text by using a sparse attention mechanism, processing longer documents more efficiently while maintaining strong performance in various NLP tasks [124]. In [46], a model combining Longformer's global and sliding window mechanisms with BERT's special classification tokens was developed. It embedded absolute and relative temporality using event tokens. They used positional indices derived from EHR record times to create a unique and shared positional encoding. A global self-attention token was used to integrate static data. It achieved an AUC of 84%.

Transformer-based NLP models were also investigated. In [39], the TAPER model, using BioBERT and a bidirectional GRU for text summarization, was proposed to obtain a unified text representation. An auto-encoder with GRUs further summarized sentence representations into a single patient text representation, aiming to reduce errors and capture crucial information effectively, achieving an AUC of 67%. In [29], multimodality analysis showed that temporal abstractions of temporal data enhanced performance, particularly with gradient inclusion, but models that used ICD-9 codes outperformed them. Overall, the ClinicalBERT model trained on clinical notes achieved the best performance with an AUC of 75%. In [125], text samples generated by the MedAug model improved the performance achieving AUCs of 79% and 82% with ClinicalBERT and MedText classifier, respectively, though they peaked with more synthesized samples before slightly declining. Transformer-based models exhibited moderate performance, with AUCs ranging from 67% to 84% and an average of 74%.

Neural Ordinary Differential Equations (ODEs) model system dynamics continuously over time using differential equations, treating time as a continuous variable, allowing for more efficient modeling of complex temporal dynamics [126]. Several methods were proposed in [30] to process time series sampled at irregular time intervals. RNN with time dynamics of code embeddings computed by neural ODEs, achieved the highest average AUC of 74%. In [127], a correlation-enhanced Multitask learning with Pearson and RNN-based Neural ODEs Model (MP-ROM) was proposed, featuring a shared bottom structure and a dynamic weighting of the loss function. Task correlation enhanced the association between sub-tasks and neural ODEs enhanced feature learning to avoid local optima. The model achieved AUCs of 74%, 74%, 74%, and 73% for predicting readmission at 5, 10, 15, and 30 days, respectively. This indicates that the 30-day prediction task benefited from multitask learning, and more improvements can be achieved by enhancing task correlations. The performance of ODE-based models was moderate, with an average AUC of 73%.

These approaches extract features without considering their medical relevance or inter-modality relationships. Recently, integrating external medical knowledge and graph methods for embedding multimodal data has shown significant performance improvements in medical applications. In [58], the Conceptual-Contextual (CC) embeddings model integrated external knowledge into text representations. Using PubMed and MIMIC-III clinical notes with BioWordVec, it retrieved context sentences based on UMLS triplets, encoded them with a bidirectional LSTM, and modeled relationships with vector addition. They achieved an AUC of 80%.

Graph Neural Networks (GNNs) [128], Graph Convolutional Networks (GCNs) [129], and Graph Attention Networks (GATs) [130] are key methods for graph classification. GNNs update node embeddings through message passing and aggregation, with attention weight learning enhancing neighbor importance. GCNs focus on aggregating features from neighboring nodes to capture local structures, while GATs use dynamic attention mechanisms to weigh neighbor importance. GCNs handle varying node degrees well, and GATs adapt to different edge significances, making both effective for analyzing complex graph-structured data. In [64], the MedText model represented clinical notes as document-level graphs, combining text and UMLS knowledge into a four-view graph to capture different interactions. It used GCN for encoding and an attention layer for decoding. the document is also encoded by a bidirectional LSTM, generating a second document-level representation, and the concatenated representations were classified using NN, achieving an AUC of 83%. In [59], ME2Vec embedded medical services, doctors, and patients, emphasizing the temporal nature of EHR data. Biased random walks [131] highlighted rare services, and GAT predicted doctor specialties using a bipartite graph. An attributed multi-graph was simplified using the duplication and annotation approach to derive patient embeddings. Training LR and RNN models with these embeddings achieved AUC scores of 59% and 60%, respectively, outperforming other embedding and matrix factorization methods.

These self-attention graph models struggle to effectively learn attention parameters from scratch, often resulting in uniformly distributed attention weights among medical concepts. To address this, [60] introduced the Graph Convolutional Transformer (GCT), which uses an attention mask and KL divergence to focus on meaningful connections and an adjacency matrix based on conditional probability for visit representations. GCT outperformed the transformer, achieving an AUC of 75% compared to 73%. Building on this, Liu, et al. [61] proposed the Statistics and Knowledge-based Graph Transformer (S_K_GT), which uses a knowledge attention network for optimized learning, achieving an AUC of 76%. In [65], the CARE-30 model used a Graph Auto-Encoder (GAE) [132] to create a Directed Acyclic Graph (DAG) [133] for capturing causal relationships among variables. Latent representations from the GAE were combined with multi-modal variables encoded by transformers, achieving an AUC of 79%. The model's interpretability was enhanced through graph weight thresholding, and robustness was improved with an average treatment effect (ATE) derived loss function [134].

Graph-based models, while effective in representing complex relationships, struggle with capturing sequential dependencies and temporal dynamics present in time series due to their inherent lack of temporal processing capabilities. To address this, in [62], a Variationally Regularized Graph Neural Network (VGNN) was introduced to enhance GNN attention. The encoder processes medical embeddings to represent the graph, while the decoder provides inferences based on the graph representations. A latent layer generated latent variables, regularized by KL divergence, approximating the distributions to Gaussian where mean and standard deviation are computed from the graph by two separate feed-forward networks achieving an AUPRC of 40%. In [66], Graph Attention and RNN-based Neural ODE Model (GROM) was proposed to convert variable-length medical code sequences into fixed-size vectors. It combined a neural ODE layer for handling irregular time series data with graph attention using CCS knowledge to learn robust diagnostic code representations, reduce noise, and achieve an AUC of 79%.

Researchers have used contrastive learning to improve graph-based models by better capturing patterns and relationships. In [63], Hypergraph Contrastive Learning (HCL) was introduced to represent complex relationships in EHR data using a Hypergraph Attention Network (HAT) [135], transformer, and GAT. HAT aggregated medical code embeddings using composer and dispatcher functions, the transformer learned code-code relationships, and GAT modeled patient-patient relationships. HCL achieved AUC scores of 72% on the eICU database and 75% on the MIMIC-III database with supervised contrastive learning, outperforming VGNN and GCT. It also reached an AUC of 69% with self-supervised learning on the eICU database, highlighting the effectiveness of contrastive learning and the importance of modeling diverse relationships in EHR data. In [67], the CodeText cross-modal Contrastive Learning (CTCL) framework tackled data heterogeneity and quality issues using a multi-view graph convolution network (MGCN) [136] and a cross-view contrastive learning module. BioClinicalBERT encoded clinical text, and a cross-modal encoder fused code and text representations, achieving an AUC of 85%. In [68], semantic annotation and Knowledge Graph (KG) embeddings were used to uniformly represent multimodal data. RDF2Vec embeddings [137] from the National Cancer Institute Thesaurus (NCIT) ontology [138] selected using the BioPortal Recommender platform [139] and a RF model achieved the best results with an AUC of 83%. The study found that multiple domain-specific ontologies did not outperform a single general-purpose ontology. Maximum performance was not achieved with discharge information alone, emphasizing the influence of data completeness, domain, and ontology appropriateness. Overall, graph and knowledge-based models showed promising performance, with AUC values ranging from 59% to 85%, and an average AUC of 75%.

Table 4 summarizes DL approaches for predicting ICU readmission, showing modest improvements over ML models. DL models fell short of expectations despite ample data availability, highlighting the problem's complexity. However, NLP significantly improved results by leveraging medical notes, and incorporating graphs and external knowledge enhanced performance. Future model advancements could potentially develop a decision-making system that meets expectations.

Fig. 2 illustrates the evolution of study methods (Statistical, ML, DL, Graph) across five periods: 2008–2013, 2014–2018, 2019–2020, 2021–2022, and 2023. Initially, from 2008–2013, only statistical methods were used (7 studies). ML methods emerged in 2014, leading to 7 studies by 2018, due to their superior performance, causing a decline in statistical methods. DL methods appeared in 2019 with 8 studies, gaining popularity for their promising results. However, over time, there was a resurgence of interest in ML models, particularly boosting techniques, due to their interpretability compared to the black-box nature of DL methods. Interpretability is crucial in healthcare applications for clinical acceptance. The current trend shows an increase in graph-based methods, with 6 studies in 2021–2022 and 3 in 2023, as they mimic the hierarchical decisions of physicians, providing a structured approach.

Table 4
DL approaches for ICU prediction.

Study	Database name (Sample size)	Time period (days) Readmission rate %	Data type				Classifier	Performance		
			Demographic	Temporal	Clinical notes	Medical codes		AUC (%)	APURC (%)	Accuracy (%)
Lin 2019 [21]	MIMIC-III (48,393)	30 (14)	🟢	🟢	🔴	🟢	LSTM+CNN	79	–	–
Zebin 2019 [38]	MIMIC-III (48,393)	30 (14)	🟢	🟢	🔴	🟢	LSTM+CNN	82	–	73
Lu 2019 [54]	MIMIC-III (48,411)	30 (14)	🔴 🔴	🔴 🔴	🟢 🔴	🔴 🟢	LSTM-CNN	79 78	48 48	75 72
Chen 2022 [121]	MIMIC-IV (67,727)	30 (14)	🟢	🟢	🔴	🟢	LSTM+CNN	63	–	65
Lovelace 2020 [49]	MIMIC-III (45,260)	30 (13) same (8)	🔴	🔴	🟢	🔴	CNN+ attention	69 71	24 17	– –
Jain 2019 [45]	MIMIC-III (34 289)	30 (22)	🔴	🔴	🟢	🔴	LSTM+ log-odds attention	71	29	–
Darabi 2020 [39]	MIMIC-III (38,597)	30	🟢	🔴	🟢	🟢	TAPER	67	68	–
Sheetrit 2023 [29]	MIMIC-III (15,424)	30 (11)	🔴 🟢	🔴 🟢	🟢 🟢	🔴 🟢	ClinclaBERT GRU	75 75	30 29	– –
Lu 2021 [125]	MIMIC-III (37 802)	30 (20)	🔴	🔴	🟢	🔴	MedAug+ MedText	82	63	–
Shickel 2022 [46]	Private (73,190)	same (6)	🟢	🟢	🔴	🔴	Longformer	84	–	–
Barbieri 2020 [30]	MIMIC-III (45,298)	30(12)	🟢	🟢	🔴	🟢	RNN (ODE time decay)	74	–	–
Niu 2023 [127]	MIMIC-III (13,383)	5 (20) 30 (41)	🟢	🟢	🔴	🟢	MP-ROM	74 73	– –	– –
Zhang 2020 [58]	MIMIC-III (48,393)	30 (14)	🔴	🔴	🟢	🔴	CC-LSTM	80	61	85
Lu 2021 [64]	MIMIC-III (48, 393)	30 (14)	🔴	🔴	🟢	🔴	MedText	83	63	–
Wu 2021 [59]	eICU (141,666)	(13)	🔴	🔴	🔴	🟢	ME2Vec+RNN ME2Vec+LR	60 59	20 19	– –
Wang 2023 [65]	MIMIC-III (38,023)	30 (16)	🔴	🟢	🟢	🟢	CARE-30	79	54	85
Choi 2020 [60]	eICU (41,026)	same (17)	🔴	🔴	🔴	🟢	GCT	75	52	–
Liu 2021 [61]	eICU (41,026)	same (17)	🔴	🔴	🔴	🟢	S_K_GT	76	–	–
Zhu 2021 [62]	eICU (41,026)	(17)	🔴	🔴	🔴	🟢	VGNN	–	40	–
Pei 2021 [66]	MIMIC-III (45,298)	30 (12)	🟢	🟢	🔴	🟢	GROM	79	–	–
Cai 2022 [63]	eICU (41,026) MIMIC-III (50,314)	(17) (21)	🔴	🔴	🔴	🟢	HCL	72 75	40 43	– –
Sun 2023 [67]	eICU (15,360)	same	🔴	🔴	🟢	🟢	CTCL	85	89	
Carvalho 2023 [68]	MIMIC-III (48,392)	30 (23)	🟢	🟢	🔴	🟢	KG embeddings +RF	83	69	–

3.3. Interpretability, model calibration and generalization

Table 5 shows the summary of studies utilizing interpretation, model calibration, and generalization.

Interpretability addresses a key issue in healthcare AI: the use of "black box" models, which lack transparency in decision-making processes. Clinicians expect to see both global feature importance and patient-specific importance. Studies using LR models clarified interpretability through model weights [26,33,57], odds ratios [37], and nomograms [34,71]. In contrast, fuzzy model studies lacked interpretability due to complex rules, and non-linearity.

ML models, except tree-based ones, were black boxes due to their intricate mathematical computations. Boosting algorithms determine feature importance based on split count reducing impurity (Gini impurity or entropy) [22,24,25,40,41]. Post hoc explanation methods such as Local Interpretable Model-agnostic Explanations (LIME) [140] or SHAP were used to demonstrate local and global interpretability [20,111]. However, post hoc methods have several shortcomings concerning robustness and adversarial attacks limiting their usefulness in health care settings [141].

Like ML models, DL and graph models are not easily interpretable due to their complex architectures, numerous parameters, and non-linear feature interactions. To address this, studies proposed using the Kolmogorov–Smirnov test [142] or dot-product attention to highlight important features [30,49]. Ablation studies have been used to

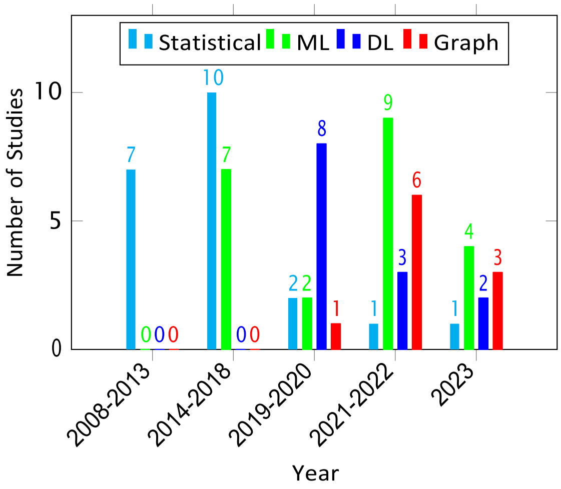

Fig. 2. Number of studies per year and method type.

address ambiguity by systematically altering features or data modalities and observing the impact on performances [21,29,39,65,66,68,113,127].

Calibration ensures that predicted probabilities align with actual outcomes, enhancing accuracy and reliability. Various calibration tests such as probability calibration curves, the Brier score [143], the Hosmer–Lemeshow goodness-of-fit test [144], the Wilcoxon signed-rank test [145], the Kruskal–Wallis test [146], post hoc sensitivity analysis [147], the Nemenyi test [148], Error bars, ANOVA test, Tukey's HSD test [149] and partial dependence plots were proposed [22,24,33,37,39–41,68,70,78]. Impact analysis studies using probability-time curves and risk thresholds, case studies and generalization to low-quality data, and the value oscillation degree are also different methods used for calibration [20,21,67]. Various studies used ablation studies to isolate and quantify the impact of components and modifications, improving predictive accuracy and validating enhancements through comparisons with original models [63–65,67,127]. Some studies explored attention function behavior to understand how these mechanisms operate within models using methods such as singular value analysis and cluster compactness [45,60,62].

Model generalization is crucial in healthcare predictive models to ensure performance on new data. External validation tests, using independent datasets, assess robustness, realistic performance, clinician trust, and weaknesses. However, few studies report external validation performance. Some observed performance drops [18,24], while others, like those validating on MIMIC-IV, noted improvements due to better data quality [41]. De Hond, et al. [110] assessed the performance of the Thoral model [20], emphasizing the need to retrain models before applying them to new data.

Implementing prediction models in production is crucial for leveraging their full potential to enhance healthcare outcomes. For a model to be practical, it must meet several key requirements. First and foremost, the model must comply with regulations regarding patient data privacy and security, ensuring that sensitive information is protected. It must consistently provide accurate and reliable predictions with high sensitivity and specificity to ensure patient safety and effective clinical outcomes. The model should be capable of handling large volumes of data and scaling with the hospital's needs, demonstrating robustness and generalization and offering timely results. Both local and global interpretability are crucial. The model should also have a user-friendly interface, regular updates, and maintenance. Lastly, the implementation and maintenance costs of the model should be proper.

Few studies implemented a practical model. INTCare is an Intelligent Decision Support System (IDSS) that published two studies [96,113]. He, et al. [17] integrated their model with a secured server and a graphical user interface (GUI) featuring input, feedback, and maintenance layers. In [101], the Intelligent ICU Patient Monitoring (IICUPM) module was proposed. Thoral, et al. [20] and De Hond, et al. [110] proposed the Pacmed Critical module.

Table 6 presents an evaluation of practical prediction models against key requirements. Red flags indicate either a lack of mention or failure to meet the criteria. Models with an AUC below 90% are deemed inadequate in terms of accuracy. Additionally, models utilizing datasets with fewer than 30,000 samples are considered unsatisfactory in terms of scalability.

4. Discussion

The analysis of the papers highlights several promising ideas for enhancing prediction models, though they need further exploration. Transfer learning [40,108] shows potential but requires validation in clinical settings. Ensemble learning [16,85] could improve accuracy but needs optimization. It is worth noting the work by Alabdulhafith et al. [150], which developed a clinical monitoring system that enhanced ICU decision-making through fog computing. Their stacking ensemble model, optimized with Genetic Algorithm (GA) and Particle Swarm Optimization (PSO), achieved impressive AUCs of 96%, 97%, and 98% with various feature sets. However, the study's focus on a limited dataset from Surgical ICUs raises concerns about the generalizability of these results to other ICU types, highlighting the need for further research to validate such systems across diverse settings. Multitask learning [127] demonstrates that the 30-day prediction task benefits from shorter-term tasks. Shickel's embedding framework [46] offers advantages by bypassing typical temporal preprocessing. Enhancing graphs with external knowledge and variational regularization may improve model generalizability. These approaches are promising but require more research to fully assess their effectiveness and practical integration.

Table 5
Summary of studies utilizing interpretation, model calibration, and generalization.

Study	Interpretation	Model calibration	Generalization
Campbell [33]	🟢	🟢	🔴
Haribhakti [70]	🔴	🟢	🔴
Jo [71]	🟢	🔴	🔴
Forst [34]	🟢	🔴	🔴
Badwi [37]	🟢	🟢	🔴
Ouanes [26]	🟢	🔴	🔴
Li [57]	🟢	🔴	🔴
Vieira [78]	🔴	🟢	🔴
He [17]	🟢	🔴	🔴
Rojas [24]	🟢	🟢	🟢
Shi [18]	🔴	🔴	🟢
Zhu [40]	🟢	🟢	🔴
Fathy [25]	🟢	🔴	🔴
Pakbin [22]	🟢	🟢	🔴
Thoral [20]	🟢	🟢	🟢
De Hond [110]	🔴	🔴	🟢
González-Nóvoa [111]	🟢	🔴	🔴
Hegselmann [41]	🟢	🟢	🟢
Veloso [113]	🟢	🔴	🔴
Lin [21]	🟢	🟢	🔴
Lovelace [49]	🟢	🔴	🔴
Jain [45]	🔴	🟢	🔴
Darabi [39]	🟢	🟢	🔴
Sheetrit [29]	🟢	🔴	🔴
Barbieri [30]	🟢	🔴	🔴
Niu [127]	🟢	🟢	🔴
Lu [64]	🔴	🟢	🔴
Wang [65]	🟢	🟢	🔴
Choi [60]	🔴	🟢	🔴
Zhu [62]	🔴	🟢	🔴
Pei [66]	🟢	🔴	🔴
Cai [63]	🔴	🟢	🔴
Sun [67]	🔴	🟢	🔴
Carvalho [68]	🟢	🟢	🔴

Table 6
Assessment of practical prediction models against key requirements.

Study	Data privacy and security	Accuracy and reliability	Scalability	Generalization	Timeliness	Interpretability	Ease of use	Maintenance and updates	Cost-Effectiveness
He [17]	🟢	🔴	🔴	🔴	🟢	🟢	🟢	🟢	🔴
Alghatani [101]	🔴	🔴	🟢	🔴	🟢	🔴	🟢	🔴	🟢
Thoral [20]	🔴	🔴	🔴	🔴	🔴	🔴	🔴	🔴	🔴

The analysis also shows that predicting ICU readmission is a challenging task for several reasons. The first challenge is the availability and complexity of public databases. Currently, only two widely used databases, MIMIC and eICU, are available. Extracting relevant data is difficult due to the lack of standardized structure, scattered information across multiple tables, and varying identifiers; requiring a manual search to identify the relevant IDs accurately. Missing or invalid data on crucial variables like patient ID and ICU admission time further complicates the process. Consequently, high-end hardware may be needed for efficient data extraction, which may not be accessible to all researchers. Detailed information on these databases is provided in Appendix A.1. The second challenge is the poor quality of ICU data, including high rates of missing values, variability in measurement units, outliers, and unreasonable values. High dimensionality and difficulty in merging data from different timestamps further contribute to biased predictions and inaccurate model performance, impacting patient outcomes. The third challenge is creating an ICU readmission predictor that addresses patient population heterogeneity, complex conditions, and imbalanced classes. Patients exhibit a wide range of conditions and comorbidities, with health status evolving during their ICU stay. Confounding factors like age, gender, and socioeconomic status also impact readmission likelihood. Additionally, the time-varying nature of risk factors and imbalanced classes (4%–14% readmissions) complicate the development of a universal prediction model.

The study highlighted a significant gap between current prediction models and practical applications. While many models are theoretically promising, they often fall short in performance, scalability, interpretability, generalizability, and clinical integration due to several factors. Variations in defining ICU readmission timeframes complicate comparisons. Studies have shifted from 2-day windows [33,37,42,108] to 30-day windows [23,27,28,57,73], influenced by research goals, healthcare contexts, and efforts to increase instances in the minority class. Some studies inaccurately treat ICU readmissions and deaths as equivalent outcomes [20,21,37,68,108,121]. However, clinical research by Krumholz, et al. [151] suggests these outcomes are orthogonal, questioning the validity of modeling them jointly. Additionally, many models are unsuitable for real-time use because they rely on data from either early [33,43,101] or end-of-stay [21,23,24,37,90,121] periods, or on discharge summaries and ICD-9 codes available only at the end of stay [44,45,49,54]. This approach ignores ongoing patient progress, resulting in a fixed readmission probability that does not reflect treatments and health changes, thus making real-time prediction impractical. Ideally, a model should utilize all available data to understand patient changes during the ICU stay for more accurate real-time predictions. Inclusion and exclusion criteria vary across different studies, variability can lead to overlooking important factors, potentially impacting the model's performance. Table 7 summarizes the inclusion and exclusion criteria for patient cohorts. Additionally, some studies may use a limited number of variables based on previous knowledge or hypotheses, potentially leaving out important predictors [16,18,20,34,37,43,82].

Many studies do not specify whether imputation methods were employed, and models frequently encounter difficulties with incomplete data. Some imputation techniques have been utilized, with the Last Observed Carried Forward (LOCF) method being the most commonly used across multiple studies [13,14,21,25,65,78,82,83,90,108,121]. MICE was applied in [23,37], KNN was used in [73], while the Expectation–Maximization (EM) algorithm [152] was used in [28,40]. Additionally, statistical techniques such as mean imputation, median imputation, interpolation, or utilizing normal variable values were employed in [18,22,29,41,46], and a weight decay term was added to the RF model in [42]. However, the impact of these imputation methods on model performance remains insufficiently studied.

Additionally, data imbalance is underexplored, with many studies using subsampling that may affect generalizability [21,24,44,73,90,91,121]. Solutions include using k-means clustering to resemble majority class [101], giving more weight to the minority class during classification [19,25,30,42], upsampling [96], or using augmentation techniques like SMOTE [17,19,38]. In [82], a balanced training set was used, with the remaining data reserved for the test set. Generating synthetic temporal data remains difficult, although methods like a teacher–student framework using Generative Pre-trained Transformer-2 (GPT-2) [153] as the teacher and CNN-LSTM as the student show promise [125]. A comprehensive evaluation of the quality of these synthetic samples and their impact on model efficiency is needed.

EHR databases offer many variables, leading to a broad range of features such as statistical, temporal, and spectral. However, this results in a large number of potentially irrelevant features. Some studies employed filter feature selection methods [154], such as Information Gain (IG) and Principal Component Analysis (PCA) in [19,73], statistical tests like chi-squared or Fisher exact test and Wilcoxon or Kruskal–Wallis test in [26,70], NMF in [40], mutual information in [25], Correlation Coefficient, Relief, and Correlation-based Feature Selection (CFS) in [90], SU in [91], and LR-Test in [37,42,71]. Other studies used wrapper methods [155], such as SFS in [16], tree search feature selection [156] in [13,78] and BFSS in [15]. Additionally, embedded methods [157] was utilized in some studies [34,41,44], while others applied ablation studies to define the important features [21]. Table 8 summarizes used preprocessing techniques. Only 39 employed at least one preprocessing technique, with only 5 studies implementing all preprocessing techniques.

Another limitation is handling inaccurate medical information, which impacts model reliability. Some studies incorrectly represented the Glasgow Coma Scale eye-opening with eight categorical values instead of the correct four [21]. Such errors arise from varied data storage techniques and a lack of medical expertise among researchers. Additionally, most studies do not address outliers or different measurement units, leading to result inconsistencies. Although some proposed pipelines handle these issues [158,159], they have not been widely adopted. Inappropriate evaluation metrics like accuracy or precision can be misleading with imbalanced data, as models may excel with the majority class but underperform with the minority [18,28,90,91,96]. Additionally, most studies lack external validation, limiting result generalizability. Many ML and DL models also suffer from poor interpretability, hindering clinical use, and few models incorporate real-time data, further restricting their practical application.

5. Future directions

The potential for predicting ICU readmission to save lives and resources is exciting, but several critical areas require attention to fully realize this potential. Interdisciplinary collaboration and diversity are essential for overcoming the challenges of current models.

To maximize ICU readmission prediction, healthcare systems should standardize EHR database design for consistent data collection, storage, and analysis and provide training for nurses to reduce entry errors. Enhancing the pipelines for data extraction from databases like MIMIC-III and eICU is crucial, requiring improvements in data preprocessing and data integration. The ideal pipeline should extract all variables in the database, not just those relevant to the specific research problem. Standardizing ICU readmission definitions and timeframes, along with consistent inclusion and exclusion criteria, is essential for accurate cohort identification and effective analysis.

Augmentation techniques generating new time series data should receive more attention due to the current lack of efficient methods. Generating realistic and meaningful synthetic EHR timeseries data requires a deep understanding of medical domain knowledge, clinical workflows, and patient health states. Furthermore, EHR data are often longitudinal, capturing a patient's medical history over time. Exploring new EHR augmentation methods, such as deep generative models and NLP, can improve EHR data quality and enhance predictive model accuracy.

Diseases are important to consider in predicting ICU readmission; however, their representation as vectors is challenging due to the complex and diverse relationships among them. Graph-based and hierarchical embedding techniques show promise in capturing disease relationships, but they can be computationally expensive and suffer

Table 7
Summary of patient cohort inclusion and exclusion criteria.

Study	Inclusion criteria				Exclusion criteria				
	Age (≥ years)	ICU LOS	1st ICU	Frequency per variable	died during ICU or hospital stay	Transferred between hospital units	Discharged from the hospital	Planned admission or discharge	Registration problems
Campbell [33]	16	–	–	–	✓	–	–	✓	–
Haribhakti [70]	18	–	–	–	✓	–	–	✓	–
Jo [71]	18	–	–	–	✓	✓	–	✓	–
Forst [34]	15	–	–	–	✓	–	✓	✓	–
Badwi [37]	16	≥4 h	–	–	✓	✓	✓	✓	–
Ouanes [26]	18	≥24 h	✓	–	✓	–	✓	✓	–
Xue [23]	–	≥12 h	✓	–	–	–	–	–	–
Li [57]	–	–	–	–	✓	–	–	✓	✓
Fialho [13]	15	≥24 h	–	All	✓	–	✓	–	–
Fialho [14]	15	≥24 h	–	≥1	✓	–	✓	–	–
Vieira [78]	15	≥24 h	–	–	✓	–	✓	–	✓
Curto [43]	–	≥24 h	–	≥2	✓	–	–	–	–
Salgado [85]	–	≥24 h	–	–	–	–	–	–	–
Viegas [16]	15	≥24 h	–	≥2	✓	–	–	–	–
Junqueira [91]	18	–	–	–	✓	–	✓	–	–
Raza [93]	18	–	✓	–	✓	✓	–	–	–
Negar [44]	–	–	–	–	–	–	–	✓	–
Silva [99]	15	–	–	–	–	–	–	–	–
He [17]	15	≥24 h	–	All	✓	–	–	–	–
Wang [42]	–	<30 days	–	–	✓	–	–	–	–
Khodadadi [102]	–	≥24 h	–	–	–	–	–	–	–
Rojas [24]	–	–	–	–	✓	✓	–	✓	–
Shi [18]	18	≥24 h	✓	–	✓	✓	✓	✓	–
Zhu [40]	16	–	–	–	–	–	–	–	–
Desautels [108]	16	≥6 h	–	≥1	–	✓	–	–	✓
Fathy [25]	18	–	✓	–	✓	–	–	–	✓
Thoral [20]	18	>30 days	–	–	–	–	✓	✓	–
De Hond [110]	18	≥12 h	–	–	–	–	–	–	–
González-Nóvoa [111]	18	–	✓	–	✓	–	✓	–	–
Hegselmann [41]	–	–	–	–	✓	–	✓	–	✓
Lin [21]	18	–	–	–	✓	–	–	–	–
Zebin [38]	18	–	–	–	✓	–	–	–	–
Lu [54]	–	–	–	–	✓	–	–	–	–
Chen [121]	18	–	–	–	✓	–	–	–	–
Darabi [39]	18	–	–	–	–	–	–	–	–
Sheetrit [29]	18	≥24 h <30 days	✓	≥1	✓	–	–	–	–
Shickel [46]	–	<10 days	–	–	–	–	–	–	–
Barbieri [30]	18	–	–	–	✓	–	–	–	–
Sun [67]	–	≥24 h	–	–	–	–	–	–	✓
Carvalho [68]	18	–	–	–	–	–	–	–	–

from overfitting. Future research should focus on developing more sophisticated disease network models that incorporate external medical knowledge and temporal relationships among diseases. Standardized disease ontologies and taxonomies can also aid in improving disease representation in prediction models.

Models that integrate time series data and medical notes are crucial as this leads to a more comprehensive understanding of a patient's condition. Time series data can provide insights into how a patient's condition changes over time and medical notes, on the other hand, provide context and detail that can help identify risk factors and underlying conditions. In addition, there is an increasing need for models that can dynamically visualize and interpret patient conditions to help doctors understand the model output and make informed decisions. Moreover, while the ultimate goal of the studies is to develop real-time applications, they have focused primarily on testing and comparing different AI models. Future research efforts should prioritize the deployment and testing of these models in real-life scenarios to ensure that they are effective and can be successfully integrated into clinical practice.

6. Conclusion

This review provides an extensive overview of 66 research papers on predicting ICU readmission using statistical, ML, and DL models. Predicting ICU readmission is challenging due to missing or invalid data, complex ICU data, and poor handling of missing or imbalanced data. To address these limitations, a standard database design, a unified definition of ICU readmission, and consistent exclusion and inclusion criteria are important. Additionally, advanced ML and DL models that integrate time series data and medical notes, along with augmentation

Table 8
Overview of preprocessing techniques utilized in reviewed studies.

Study	Feature selection	Imputation technique	Rebalance technique
Haribhakti [70]	🟢	🔴	🔴
Jo [71]	🟢	🔴	🔴
Forst [34]	🟢	🔴	🔴
Badwi [37]	🟢	🟢	🔴
Ouanes [26]	🟢	🔴	🔴
Xue [23]	🔴	🟢	🔴
Moerschbacher [73]	🟢	🟢	🟢
Fialho [13]	🟢	🟢	🔴
Fialho [14]	🔴	🟢	🔴
Vieira [78]	🟢	🟢	🔴
Sargo [15]	🟢	🔴	🔴
Ferreira [82]	🔴	🟢	🟢
Salgado [83]	🔴	🟢	🔴
Viegas [16]	🟢	🔴	🔴
Venugopalan [28]	🔴	🟢	🔴
Inan [90]	🟢	🟢	🟢
Junqueira [91]	🟢	🔴	🟢
Negar [44]	🟢	🔴	🟢
Braga [96]	🔴	🔴	🟢
He [17]	🔴	🔴	🟢
Wang [42]	🟢	🟢	🟢
Alghatani [101]	🔴	🔴	🟢
Rojas [24]	🔴	🔴	🟢
Shi [18]	🔴	🟢	🔴
Zhu [40]	🟢	🟢	🔴
Loreto [19]	🟢	🟢	🟢
Desautels [108]	🔴	🟢	🔴
Fathy [25]	🟢	🟢	🟢
Pakbin [22]	🔴	🟢	🔴
Thoral [20]	🟢	🔴	🔴
Hegselmann [41]	🟢	🟢	🔴
Lin [21]	🔴	🟢	🟢
Zebin [38]	🔴	🔴	🟢
Chen [121]	🔴	🟢	🟢
Sheetrit [29]	🔴	🟢	🔴
Lu [125]	🔴	🔴	🟢
Shickel [46]	🔴	🟢	🔴
Barbieri [30]	🔴	🔴	🟢
Wang 2023 [65]	🔴	🟢	🔴

techniques and powerful disease embedding, can improve the accuracy and reliability of prediction models.

CRediT authorship contribution statement

Waleed Fathy: Conceptualization, Data curation, Formal analysis, Methodology, Resources, Validation, Visualization, Writing – original draft, Writing – review & editing. **Guillaume Emeriaud:** Data curation, Funding acquisition, Supervision, Validation. **Farida Cheriet:** Funding acquisition, Project administration, Supervision, Validation.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Guillaume Emeriaud reports financial support was provided by Fonds FSISSS of MedTeq. Waleed Fathy reports financial support was provided by Opsidian Create program of Natural Science and Engineering Research Council of Canada.

Acknowledgments

This work is financially supported by Opsidian Create program of Natural Science and Engineering Research Council of Canada and the Fonds FSISSS of Medteq.

Appendix. Database

A.1. Overview of used databases

Ensuring replicable methodology is essential in data science and typically requires data sharing. However, the medical and clinical fields face ethical limitations due to the sensitivity and confidentiality of patient data. Balancing these ethical concerns with the need for reproducibility highlights the importance of open-access datasets in medical and clinical research.

The outcomes analysis reveals a predominant reliance on publicly available databases, notably MIMIC. The MIMIC databases and the eICU Collaborative Research Database are key public datasets for critical care research. These datasets serve as invaluable resources for advancing critical care research. MIMIC provides comprehensive clinical data, including physiological waveforms, demographics, and laboratory results. Three versions of the MIMIC series have been used: MIMIC-II (2001–2008) [17,23,40,83], MIMIC-III (2001–2012) [57,65,68,73], and MIMIC-IV (up to 2019) [18,41,121]. These versions progressively improve data quality and expand critical care research insights. The eICU database, sourced from various institutions, offers detailed ICU patient information, spanning demographics, vital signs, laboratory results, and outcomes, ensuring a diverse representation. The eICU database has been used in several studies [37,63,67,102].

In addition, various private databases have been utilized in studies, including Scottish ICU database [33], Seoul National University Bundang Hospita [71], Liverpool hospital in Australia [34], French Outcomerea network [26], Hospital da Luz in Portugal [14], Centro Hospitalar do Porto [96,113], Anhui hospital [42], University of Chicago Medical Center [24], Stanford Medicine Research Data Repository (STARR) [18], Children's Healthcare of Atlanta (CHOA) [40], Brazilian university hospital [19], Cambridge University Hospital (CUH) [108], Amsterdam UMC [20,110], Leiden university medical centre [110], a tertiary care hospital in the UK [108], an academic hospital in the United States [24], a Brazilian university hospital [19], Anesthesiology, Intensive Care and Pain Medicine at the University Hospital Münster (ANIT-UKM) [41], and the University of Florida Institutional [46].

A.2. Used data types

EHR databases store a wide array of patient information, including demographic details, vital signs, laboratory results, interventions, medical history, diagnosis, medications, procedures, treatment, imaging data, and clinical notes. These data types can be classified into five main groups:

1. Demographic Data: Include information like age, gender, and comorbidity, providing baseline characteristics for patients.
2. Temporal data: Comprise vital signs, interventions, laboratory tests, and time series data of treatments, offering dynamic health tracking over time.
3. Medical Codes: Represent diagnosis, procedures, and medication codes, providing insight into the medical history and treatment of patients.
4. Text Data: Include clinical notes, particularly discharge summaries, and other textual information, providing qualitative information on patient care.
5. Images: Comprise diagnostic images such as X-rays and Magnetic Resonance Imaging (MRI), providing visual information for diagnostics.

Each data type plays a crucial role in understanding the health status and history of a patient. Traditionally, ICU readmission prediction predominantly relied on demographics, temporal data, and comorbidities. Clinical notes are increasingly used to supplement these sources.

Current trends favor a comprehensive approach, utilizing various data modalities, including medical codes, for a holistic view of patient health and readmission risks. Interestingly, medical images remain largely unexplored for ICU readmission prediction, suggesting a potential area for future research. Effective representation of these data is crucial for accurate prediction models in healthcare.

The researchers selected variables for their models based on various approaches. Some chose variables according to previous medical knowledge, focusing on factors known to be relevant in predicting ICU readmission or following previous studies or guidelines, incorporating variables that are significant in similar research [18,20,34,43]. Additionally, some studies included all allowed variables with low missing rates and presence for the majority of patients, aiming to capture a comprehensive range of factors that could impact patient outcomes [28,33,39,41].

References

[1] Rosenberg AL, Watts C. Patients readmitted to ICUs: a systematic review of risk factors and outcomes. Chest 2000;118:492–502. http://dx.doi.org/10.1378/CHEST.118.2.492.

[2] Bardak B, Tan M. Improving clinical outcome predictions using convolution over medical entities with multimodal learning. Artif Intell Med 2021;117:102112. http://dx.doi.org/10.1016/J.ARTMED.2021.102112.

[3] Markazi-Moghaddam N, Fathi M, Ramezankhani A. Risk prediction models for intensive care unit readmission: A systematic review of methodology and applicability. Aust Crit Care 2020;33:367–74. http://dx.doi.org/10.1016/J.AUCC.2019.05.005.

[4] Syed M, et al. Application of machine learning in intensive care unit (ICU) settings using MIMIC dataset: Systematic review. Informatics 2021;8. http://dx.doi.org/10.3390/INFORMATICS8010016.

[5] Teo K, et al. Current trends in readmission prediction: An overview of approaches. Arab J Sci Eng 2021;1–18. http://dx.doi.org/10.1007/s13369-021-06040-5.

[6] Herland M, Khoshgoftaar TM, Wald R. Survey of clinical data mining applications on big data in health informatics. In: 12th international conference on machine learning and applications. Vol. 2, IEEE Computer Society; 2013, p. 465–72. http://dx.doi.org/10.1109/ICMLA.2013.163.

[7] Herland M, Khoshgoftaar TM, Wald R. A review of data mining using big data in health informatics. J Big Data 2014;1:1–35. http://dx.doi.org/10.1186/2196-1115-1-2.

[8] Sharma A, et al. Mortality prediction of icu patients using machine leaning: A survey. Proc Int Conf Comput Data Anal 2017;Part F130280:49–53. http://dx.doi.org/10.1145/3093241.3093267.

[9] Ruppert MM, et al. Predictive modeling for readmission to intensive care: A systematic review. Crit Care Explor 2023;5:E0848. http://dx.doi.org/10.1097/CCE.0000000000000848.

[10] Ng MY, et al. The AI life cycle: A holistic approach to creating ethical AI for health decisions. Nature Med 2022;28. http://dx.doi.org/10.1038/s41591-022-01993-y.

[11] Page MJ, et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ 2021;372. http://dx.doi.org/10.1136/bmj.n71.

[12] Moons KG, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: The CHARMS checklist. PLOS Med 2014;11:e1001744. http://dx.doi.org/10.1371/JOURNAL.PMED.1001744.

[13] Fialho AS, et al. Data mining using clinical physiology at discharge to predict icu readmissions. Expert Syst Appl 2012;39(18):13158–65. http://dx.doi.org/10.1016/J.ESWA.2012.05.086.

[14] Fialho AS, et al. Predicting intensive care unit readmissions using probabilistic fuzzy systems. IEEE Int Conf Fuzzy Syst (FUZZ-IEEE) 2013. http://dx.doi.org/10.1109/FUZZ-IEEE.2013.6622414.

[15] Sargo JAG, et al. Binary fish school search applied to feature selection: Application to icu readmissions. IEEE Int Conf Fuzzy Syst (FUZZ-IEEE) 2014;1366–73. http://dx.doi.org/10.1109/FUZZ-IEEE.2014.6891802.

[16] Viegas R, et al. Daily prediction of icu readmissions using feature engineering and ensemble fuzzy modeling. Expert Syst Appl 2017;79:244–53. http://dx.doi.org/10.1016/J.ESWA.2017.02.036.

[17] He L, Wang H, Rezaeiahari M, Chou CA. An embedded machine learning model for early detection and intervention of high-risk intensive care unit readmission patients. In: Proceedings - 2022 IEEE international conference on bioinformatics and biomedicine. BIBM 2022, Institute of Electrical and Electronics Engineers Inc.; 2022, p. 1544–9. http://dx.doi.org/10.1109/BIBM55620.2022.9995664.

[18] Shi K, et al. Predicting unplanned 7-day intensive care unit readmissions with machine learning models for improved discharge risk assessment. AMIA Annu Symp Proc 2022;2022:446.

[19] Loreto M, Lisboa T, Moreira VP. Early prediction of ICU readmissions using classification algorithms. Comput Biol Med 2020;118:103636. http://dx.doi.org/10.1016/J.COMPBIOMED.2020.103636.

[20] Thoral PJ, et al. Explainable machine learning on AmsterdamUMCdb for ICU discharge decision support: Uniting intensivists and data scientists. Crit Care Explor 2021;3:e0529. http://dx.doi.org/10.1097/CCE.0000000000000529.

[21] Lin YW, et al. Analysis and prediction of unplanned intensive care unit readmission using recurrent neural networks with long short-term memory. PLOS ONE 2019;14(7):e0218942. http://dx.doi.org/10.1371/JOURNAL.PONE.0218942.

[22] Pakbin A, et al. Prediction of icu readmissions using data at patient discharge. In: 40th annual international conference of the IEEE engineering in medicine and biology society. EMBC, Vol. 2018, Annu Int Conf IEEE Eng Med Biol Soc; 2018, p. 4932–5. http://dx.doi.org/10.1109/EMBC.2018.8513181.

[23] Xue Y, Klabjan D, Luo Y. Predicting icu readmission using grouped physiological and medication trends. Artif Intell Med 2019;95:27–37. http://dx.doi.org/10.1016/J.ARTMED.2018.08.004.

[24] Rojas JC, et al. Predicting intensive care unit readmission with machine learning using electronic health record data. Ann Am Thorac Soc 2018;15(7):846–53. http://dx.doi.org/10.1513/AnnalsATS.201710-787OC.

[25] Fathy W, Emeriaud G, Cheriet F. Prediction of icu readmission using LightGBM classifier. In: 2023 IEEE 20th international symposium on biomedical imaging. ISBI, 2023, p. 1–4. http://dx.doi.org/10.1109/ISBI53787.2023.10230835.

[26] Ouanes I, et al. A model to predict short-term death or readmission after intensive care unit discharge. J Crit Care 2012;27:422.e1–9. http://dx.doi.org/10.1016/J.JCRC.2011.08.003.

[27] Caballero K, Akella R. Dynamic estimation of the probability of patient readmission to the icu using electronic medical records. AMIA Annu Symp Proc 2015;2015:1831.

[28] Venugopalan J, et al. Combination of static and temporal data analysis to predict mortality and readmission in the intensive care. In: 39th annual international conference of the IEEE engineering in medicine and biology society. EMBC, Vol. 2017, NIH Public Access; 2017, p. 2570. http://dx.doi.org/10.1109/EMBC.2017.8037382.

[29] Sheetrit E, Brief M, Elisha O. Predicting unplanned readmissions in the intensive care unit: a multimodality evaluation. Sci Rep 2023;13:1–9. http://dx.doi.org/10.1038/s41598-023-42372-y, 1.

[30] Barbieri S, et al. Benchmarking deep learning architectures for predicting readmission to the ICU and describing patients-at-risk. Sci Rep 2020;10(1):1–10. http://dx.doi.org/10.1038/s41598-020-58053-z.

[31] Knaus WA, et al. APACHE—acute physiology and chronic health evaluation: a physiologically based classification system. Crit Care Med 1981;9:591–7.

[32] Vincent JL, et al. The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. Intensive Care Med 1996;22:707–10. http://dx.doi.org/10.1007/BF01709751.

[33] Campbell AJ, et al. Predicting death and readmission after intensive care discharge. Br J Anaesth 2008;100:656–62. http://dx.doi.org/10.1093/BJA/AEN069.

[34] Frost SA, et al. Readmission to intensive care: development of a nomogram for individualising risk. Crit Care Resusc 2010.

[35] Choi Y, Chiu CMY-I, Sontag D. Learning low-dimensional representations of medical concepts. AMIA Summits Transl Sci Proc 2016;2016:41.

[36] Clinical Classifications Software (CCS) for ICD-9-CM. URL https://hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp.

[37] Badawi O, Breslow MJ. Readmissions and death after ICU discharge: Development and validation of two predictive models. PLOS ONE 2012;7:e48758. http://dx.doi.org/10.1371/JOURNAL.PONE.0048758.

[38] Zebin T, Chaussalet TJ. Design and implementation of a deep recurrent model for prediction of readmission in urgent care using electronic health records. EEE Conf Comput Intell Bioinform Comput Biol (CIBCB) 2019. http://dx.doi.org/10.1109/CIBCB.2019.8791466.

[39] Darabi S, et al. TAPER: Time-aware patient EHR representation. IEEE J Biomed Heal Inform 2019;24(11):3268–75. http://dx.doi.org/10.1109/JBHI.2020.2984931.

[40] Zhu Y, et al. Domain adaptation using convolutional autoencoder and gradient boosting for adverse events prediction in the intensive care unit. Front Artif Intell 2022;5. http://dx.doi.org/10.3389/FRAI.2022.640926.

[41] Hegselmann S, et al. Development and validation of an interpretable 3 day intensive care unit readmission prediction model using explainable boosting machines. Front Med 2022;9. http://dx.doi.org/10.3389/fmed.2022.960296.

[42] Wang B, et al. Predictive classification of ICU readmission using weight decay random forest. Future Gener Comput Syst 2021;124:351–60. http://dx.doi.org/10.1016/J.FUTURE.2021.06.011.

[43] Curto S, et al. Predicting ICU readmissions based on bedside medical text notes. IEEE Int Conf Fuzzy Syst (FUZZ-IEEE) 2016;2144–51. http://dx.doi.org/10.1109/FUZZ-IEEE.2016.7737956.

[44] Orangi-Fard N, Akhbardeh A, Sagreiya H. Predictive model for ICU readmission based on discharge summaries using machine learning and natural language processing. Informatics 2022;9(1):10. http://dx.doi.org/10.3390/INFORMATICS9010010.

[45] Jain S, Mohammadi R, Wallace BC. An analysis of attention over clinical notes for predictive tasks. In: Proceedings of the 2nd clinical natural language processing workshop. 2019, p. 15–21. http://dx.doi.org/10.18653/v1/W19-1902.

[46] Shickel B, et al. Multi-dimensional patient acuity estimation with longitudinal EHR tokenization and flexible transformer networks. Front Digit Heal 2022;4. http://dx.doi.org/10.3389/FDGTH.2022.1029191.

[47] Mikolov T, et al. Efficient estimation of word representations in vector space. In: International conference on learning representations. 2013.

[48] Zhang Y, et al. BioWordVec, improving biomedical word embeddings with subword information and MeSH. Sci Data 2019;6:1–9. http://dx.doi.org/10.1038/s41597-019-0055-0.

[49] Lovelace J, et al. Dynamically extracting outcome-specific problem lists from clinical notes with guided multi-headed attention. 2020, p. 245–70.

[50] Devlin J, et al. BERT: Pre-training of deep bidirectional transformers for language understanding. In: North American chapter of the association for computational linguistics. Vol. 1, 2019, p. 4171–86.

[51] Lee J, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2019;36(4):1234–40. http://dx.doi.org/10.1093/bioinformatics/btz682.

[52] Huang K, Altosaar J, Ranganath R. ClinicalBERT: Modeling clinical notes and predicting hospital readmission. 2019, ArXiv abs/1904.05342.

[53] Nickel M, Kiela D. Poincaré embeddings for learning hierarchical representations. In: Advances in neural information processing systems. Vol. 30, Curran Associates, Inc.; 2017.

[54] Lu Q, et al. Learning electronic health records through hyperbolic embedding of medical ontologies. In: Proceedings of the 10th ACM international conference on bioinformatics, computational biology and health informatics. BCB '19, New York, NY, USA: Association for Computing Machinery; 2019, p. 338–46. http://dx.doi.org/10.1145/3307339.3342148.

[55] Humphreys BL, et al. The unified medical language system: An informatics research collaboration. J Am Med Inform Assoc 1998;5:1–11. http://dx.doi.org/10.1136/JAMIA.1998.0050001.

[56] SNOMED International. URL https://www.snomed.org/.

[57] Li Z, et al. Early prediction of 30-day ICU re-admissions using natural language processing and machine learning. Biomed Stat Inform 2019;4(3):22. http://dx.doi.org/10.11648/j.bsi.20190403.11, arXiv:1910.02545.

[58] Zhang X, Dou D, Wu J. Learning conceptual-contextual embeddings for medical text. AAAI 2020;34(05):9579–86. http://dx.doi.org/10.1609/AAAI.V34I05.6504.

[59] Wu T, et al. Leveraging graph-based hierarchical medical entity embedding for healthcare applications. Sci Rep 2021;11(1):1–13. http://dx.doi.org/10.1038/s41598-021-85255-w.

[60] Choi E, et al. Learning the graphical structure of electronic health records with graph convolutional transformer. AAAI Conf Artif Intell 2020;606–13. http://dx.doi.org/10.1609/aaai.v34i01.5400.

[61] Liu X, et al. Research on intelligent diagnosis model of electronic medical record based on graph transformer. In: 6th international conference on computational intelligence and applications. ICCIA, 2021, p. 73–8. http://dx.doi.org/10.1109/ICCIA52886.2021.00022.

[62] Zhu W, Razavian N. Variationally regularized graph-based representation learning for electronic health records. In: Proceedings of the conference on health, inference, and learning. New York, NY, USA: Association for Computing Machinery; 2021, p. 1–13. http://dx.doi.org/10.1145/3450439.3451855.

[63] Cai D, et al. Hypergraph contrastive learning for electronic health records. In: Proceedings of the 2022 SIAM international conference on data mining. SDM, Society for Industrial and Applied Mathematics Publications; 2022, p. 127–35. http://dx.doi.org/10.1137/1.9781611977172.15.

[64] Lu Q, Nguyen TH, Dou D. Predicting patient readmission risk from medical text via knowledge graph enhanced multiview graph convolution. In: Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval. 2021, p. 1990–4. http://dx.doi.org/10.1145/3404835.3463062.

[65] Wang L, et al. CARE-30: A causally driven multi-modal model for enhanced 30-day ICU readmission predictions. IEEE Int Conf Bioinform Biomed (BIBM) 2023;1509–16. http://dx.doi.org/10.1109/BIBM58861.2023.10385349.

[66] Pei S, et al. Readmission prediction with knowledge graph attention and RNN-based ordinary differential equations. Knowl Sci Eng Manag 2021;12817 LNAI:559–70. http://dx.doi.org/10.1007/978-3-030-82153-1_46.

[67] Sun M, et al. A cross-modal clinical prediction system for intensive care unit patient outcome. Knowl-Based Syst 2024;283:111160. http://dx.doi.org/10.1016/J.KNOSYS.2023.111160.

[68] Carvalho RMS, Oliveira D, Pesquita C. Knowledge graph embeddings for ICU readmission prediction. BMC Med Inform Decis Mak 2023;23:1–17. http://dx.doi.org/10.1186/s12911-022-02070-7.

[69] Oord AV, Li Y, Vinyals O. Representation learning with contrastive predictive coding. 2018, http://dx.doi.org/10.48550/arXiv.1807.03748.

[70] Haribhakti N, et al. A simple scoring tool to predict medical intensive care unit readmissions based on both patient and process factors. J Gen Intern Med 2021;36:901–7. http://dx.doi.org/10.1007/S11606-020-06572-W/FIGURES/2.

[71] Jo YS, et al. Readmission to medical intensive care units: Risk factors and prediction. Yonsei Med J 2015;56:543. http://dx.doi.org/10.3349/YMJ.2015.56.2.543.

[72] Azur MJ, et al. Multiple imputation by chained equations: what is it and how does it work? Int J Methods Psychiatr Res 2011;20:40. http://dx.doi.org/10.1002/MPR.329.

[73] Moerschbacher A, He Z. Building prediction models for 30-day readmissions among ICU patients using both structured and unstructured data in electronic health records. IEEE Int Conf Bioinform Biomed (BIBM) 2023;4368–73. http://dx.doi.org/10.1109/BIBM58861.2023.10385612.

[74] Takagi T, Sugeno M. Fuzzy identification of systems and its applications to modeling and control. IEEE Trans Syst Man Cybern 1985;SMC-15:116–32. http://dx.doi.org/10.1109/TSMC.1985.6313399.

[75] Bezdek JC. Pattern recognition with fuzzy objective function algorithms. Adv Appl Pattern Recognit 1981. http://dx.doi.org/10.1007/978-1-4757-0450-1.

[76] Izakian H, Pedrycz W, Jamal I. Clustering spatiotemporal data: An augmented fuzzy C-means. IEEE Trans Fuzzy Syst 2013;21:855–68. http://dx.doi.org/10.1109/TFUZZ.2012.2233479.

[77] Berg JVD, Kaymak U, Bergh WMVD. Fuzzy classification using probability-based rule weighting. IEEE Int Conf Fuzzy Syst 2002;2:991–6. http://dx.doi.org/10.1109/FUZZ.2002.1006639.

[78] Vieira SM, et al. A decision support system for ICU readmissions prevention. Jt IFSA World Congr NAFIPS Annu Meet (IFSA/NAFIPS) 2013;251–6. http://dx.doi.org/10.1109/IFSA-NAFIPS.2013.6608408.

[79] Homem N, Carvalho JP. Authorship identification and author fuzzy "fingerprints". In: Annual meeting of the North American fuzzy information processing society. 2011, p. 1–6. http://dx.doi.org/10.1109/NAFIPS.2011.5751998.

[80] Chen C-H, Hong T-P, Tseng VS. Finding Pareto-front membership functions in fuzzy data mining. Int J Comput Intell Syst 2012;5(2):343–54. http://dx.doi.org/10.1080/18756891.2012.685314.

[81] Filho CJ, et al. A novel search algorithm based on fish school behavior. IEEE Int Conf Syst Man Cybern 2008;2646–51. http://dx.doi.org/10.1109/ICSMC.2008.4811695.

[82] Ferreira MC, et al. Fuzzy modeling based on Mixed Fuzzy Clustering for health care applications. IEEE Int Conf Fuzzy Syst (FUZZ-IEEE) 2015;2015-Novem. http://dx.doi.org/10.1109/FUZZ-IEEE.2015.7338028.

[83] Salgado CM, Vieira SM, Sousa JM. Fuzzy modeling based on mixed fuzzy clustering for multivariate time series of unequal lengths. Commun Comput Inf Sci 2016;611:741–51. http://dx.doi.org/10.1007/978-3-319-40581-0_60.

[84] Fernandes MP, et al. Multimodeling for the prediction of patient readmissions in Intensive Care Units. IEEE Int Conf Fuzzy Syst (FUZZ-IEEE) 2014;1837–42. http://dx.doi.org/10.1109/FUZZ-IEEE.2014.6891779.

[85] Salgado CM, et al. Ensemble fuzzy classifiers design using weighted aggregation criteria. IEEE Int Conf Fuzzy Syst (FUZZ-IEEE) 2015;2015-Novem. http://dx.doi.org/10.1109/FUZZ-IEEE.2015.7338110.

[86] Gustafson DE, Kessel WC. Fuzzy clustering with a fuzzy covariance matrix. IEEE Conf Decis Control 1978;761–6. http://dx.doi.org/10.1109/CDC.1978.268028.

[87] West M, Harrison J. The dynamic linear model. In: Bayesian forecasting and dynamic models. New York, NY: Springer; 1989, p. 105–41. http://dx.doi.org/10.1007/978-1-4757-9365-9_4.

[88] Lafferty JD, McCallum A, Pereira F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: International conference on machine learning. 2001.

[89] Yang ZR, Yang Z. Artificial neural networks. Compr Biomed Phys 2014;6:1–17. http://dx.doi.org/10.1016/B978-0-444-53632-7.01101-1.

[90] Inan TT, et al. A decision support model to predict ICU readmission through data mining approach. In: PACIS 2018 proceedings. 2018.

[91] Junqueira AR, Mirza F, Baig MM. A machine learning model for predicting ICU readmissions and key risk factors: analysis from a longitudinal health records. Heal Technol 2019;9(3):297–309. http://dx.doi.org/10.1007/s12553-019-00329-0.

[92] Zhang L, Chen X. Feature selection methods based on symmetric uncertainty coefficients and independent classification information. IEEE Access 2021;9:13845–56. http://dx.doi.org/10.1109/ACCESS.2021.3049815.

[93] Raza S, Bashir SR. Auditing ICU readmission rates in an clinical database: An analysis of risk factors and clinical outcomes. IEEE Int Conf Heal Inform 2023;722–6. http://dx.doi.org/10.1109/ICHI57859.2023.00132.

[94] Boser BE, Guyon IM, Vapnik VN. A training algorithm for optimal margin classifiers. In: Proceedings of the fifth annual ACM workshop on computational learning theory. Publ by ACM; 1992, p. 144–52. http://dx.doi.org/10.1145/130385.130401.

[95] Webb GI. Naïve Bayes. Encycl Mach Learn 2011;713–4. http://dx.doi.org/10.1007/978-0-387-30164-8_576.

[96] Braga P, et al. Data mining models to predict patient's readmission in intensive care units. In: ICAART 2014 - Proceedings of the 6th international conference on agents and artificial intelligence. Vol. 1, SciTePress; 2014, p. 604–10. http://dx.doi.org/10.5220/0004907806040610.

[97] Quinlan JR. Induction of decision trees. Mach Learn 1986;1:81–106. http://dx.doi.org/10.1007/BF00116251.

[98] Breiman L. Random forests. Mach Learn 2001;45:5–32. http://dx.doi.org/10.1023/A:1010933404324.

[99] Silva C, Vieira SM, Sousa JM. Fuzzy decision tree to predict readmissions in intensive care unit. Lect Notes Electr Eng 2015;321 LNEE:365–73. http://dx.doi.org/10.1007/978-3-319-10380-8_35.

[100] Liu X, Pedrycz W. The development of fuzzy decision trees in the framework of Axiomatic Fuzzy Set logic. Appl Soft Comput J 2007;7:325–42. http://dx.doi.org/10.1016/J.ASOC.2005.07.003.

[101] Alghatani K, et al. Precision clinical medicine through machine learning: Using high and low quantile ranges of vital signs for risk stratification of ICU patients. IEEE Access 2022;10:52418–30. http://dx.doi.org/10.1109/ACCESS.2022.3175304.

[102] Khodadadi A, et al. Improving diagnostics with deep forest applied to electronic health records. Sensors 2023;23. http://dx.doi.org/10.3390/S23146571.

[103] Freund Y, Schapire RE. A desicion-theoretic generalization of on-line learning and an application to boosting. Lecture Notes in Comput Sci 1995;904:23–37. http://dx.doi.org/10.1007/3-540-59119-2_166.

[104] Friedman JH. Greedy function approximation: A gradient boosting machine. Ann Statist 2001;29:1189–232. http://dx.doi.org/10.1214/AOS/1013203451.

[105] Chen T, Guestrin C. XGBoost: A scalable tree boosting system. ACM SIGKDD Int Conf Knowl Discov Data Min 2016;13-17-August-2016:785–94. http://dx.doi.org/10.1145/2939672.2939785.

[106] Ke G, et al. LightGBM: A highly efficient gradient boosting decision tree. Adv Neural Inf Process Syst 2017;30.

[107] Chawla NV, et al. SMOTE: Synthetic minority over-sampling technique. Artif Intell Res 2011;16:321–57. http://dx.doi.org/10.1613/jair.953.

[108] Desautels T, et al. Prediction of early unplanned intensive care unit readmission in a UK tertiary care hospital: a cross-sectional machine learning approach. BMJ Open 2017;7(9):e017199. http://dx.doi.org/10.1136/bmjopen-2017-017199.

[109] Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Int Conf Neural Inf Process Syst 2017;2017-December:4766–75.

[110] Hond AAD, et al. Predicting readmission or death after discharge from the ICU: External validation and retraining of a machine learning model. Crit Care Med 2023;51:291–300. http://dx.doi.org/10.1097/CCM.0000000000005758.

[111] González-Nóvoa JA, et al. Improving intensive care unit early readmission prediction using optimized and explainable machine learning. Int J Environ Res Public Heal 2023;20. http://dx.doi.org/10.3390/IJERPH20043455.

[112] Watanabe S. Tree-structured parzen estimator: Understanding its algorithm components and their roles for better empirical performance. 2023, arXiv preprint arXiv:2304.11127.

[113] Veloso R, et al. A clustering approach for predicting readmissions in intensive medicine. Procedia Technol 2014;16:1307–16. http://dx.doi.org/10.1016/J.PROTCY.2014.10.147.

[114] Jin X, Han J. K-means clustering. Encycl Mach Learn 2011;563–4. http://dx.doi.org/10.1007/978-0-387-30164-8_425.

[115] Park HS, Jun CH. A simple and fast algorithm for K-medoids clustering. Expert Syst Appl 2009;36:3336–41. http://dx.doi.org/10.1016/J.ESWA.2008.01.039.

[116] Pelleg D, Moore AW. X-means: Extending K-means with efficient estimation of the number of clusters. In: Proceedings of the seventeenth international conference on machine learning. ICML '00, San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.; 2000, p. 727–34.

[117] Sherstinsky A. Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network. Phys D: Nonlinear Phenom 2020;404:132306. http://dx.doi.org/10.1016/J.PHYSD.2019.132306.

[118] Teuwen J, Moriakov N. Convolutional neural networks. Handb Med Image Comput Comput Assist Interv 2020;481–501. http://dx.doi.org/10.1016/B978-0-12-816176-0.00025-9.

[119] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997;9:1735–80. http://dx.doi.org/10.1162/NECO.1997.9.8.1735.

[120] Chung J, et al. Empirical evaluation of gated recurrent neural networks on sequence modeling. NIPS Work Deep Learn 2014.

[121] Chen Q, et al. Outcome-oriented predictive process monitoring to predict unplanned ICU readmission in MIMIC-IV database. ECIS Res-in-Prog Pap 2022.

[122] Vaswani A, et al. Attention is all you need. Neural Inf Process Syst (NIPS) 2017;5999–6009.

[123] Denny JC, et al. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations. Bioinformatics 2010;26:1205–10. http://dx.doi.org/10.1093/BIOINFORMATICS/BTQ126.

[124] Beltagy I, Peters ME, Cohan A. Longformer: The long-document transformer. 2020, http://dx.doi.org/10.48550/arXiv.2004.05150, Arxiv.

[125] Lu Q, Dou D, Nguyen TH. Textual data augmentation for patient outcomes prediction. IEEE Int Conf Bioinform Biomed (BIBM) 2021;2817–21. http://dx.doi.org/10.1109/BIBM52615.2021.9669861.

[126] Chen RTQ, et al. Neural ordinary differential equations. Neural Inf Process Syst (NIPS) 2018;109:31–60.

[127] Niu K, et al. Intensive Care Unit readmission prediction with correlation enhanced multi-task learning. Comput Electr Eng 2023;110:108780. http://dx.doi.org/10.1016/J.COMPELECENG.2023.108780.

[128] Zhou J, et al. Graph neural networks: A review of methods and applications. AI Open 2020;1:57–81. http://dx.doi.org/10.1016/J.AIOPEN.2021.01.001.

[129] Zhang S, et al. Graph convolutional networks: a comprehensive review. Comput Soc Netw 2019;6:1–23. http://dx.doi.org/10.1186/s40649-019-0069-y.

[130] Veličković P, et al. Graph attention networks. Int Conf Learn Represent (ICLR) 2017. http://dx.doi.org/10.17863/CAM.48429.

[131] Leleux P, et al. Design of biased random walks on a graph with application to collaborative recommendation. Phys A 2022;590:126752. http://dx.doi.org/10.1016/J.PHYSA.2021.126752.

[132] Ng I, et al. A graph autoencoder approach to causal structure learning. 2019, http://dx.doi.org/10.48550/arXiv.1911.07420, ArXiv.

[133] Foraita R, Spallek J, Zeeb H. Directed acyclic graphs. Handb Epidemiology: Second Ed 2014;1481–517. http://dx.doi.org/10.1007/978-0-387-09834-0_65.

[134] Kuang K, et al. Causal inference. Engineering 2020;6:253–63. http://dx.doi.org/10.1016/J.ENG.2019.08.016.

[135] Chen C, et al. Hypergraph attention networks. IEEE Int Conf Trust Secur Priv Comput Commun (TrustCom) 2020;1560–5. http://dx.doi.org/10.1109/TRUSTCOM50675.2020.00215.

[136] Khan MR, Blumenstock JE. Multi-GCN: Graph convolutional networks for multi-view networks, with applications to global poverty. AAAI Conf Artif Intell 2019;33:606–13. http://dx.doi.org/10.1609/AAAI.V33I01.3301606.

[137] Ristoski P, et al. RDF2Vec: RDF graph embeddings and their applications. Semant Web 2019;10:721–52. http://dx.doi.org/10.3233/SW-180317.

[138] NCI Thesaurus, URL https://ncit.nci.nih.gov/ncitbrowser/.

[139] Martínez-Romero M, et al. NCBO Ontology Recommender 2.0: An enhanced approach for biomedical ontology recommendation. J Biomed Semant 2017;8:1–22. http://dx.doi.org/10.1186/S13326-017-0128-Y.

[140] Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?": Explaining the predictions of any classifier. Conf North Am Chapter Comput Linguist: Demonstr 2016;97–101. http://dx.doi.org/10.18653/v1/n16-3020.

[141] Slack D, et al. Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. AAAI/ACM Conf AI Ethics Soc 2019;180–6. http://dx.doi.org/10.1145/3375627.3375830.

[142] Massey FJ. The Kolmogorov-Smirnov test for goodness of fit. Am Stat Assoc 1951;46:68. http://dx.doi.org/10.2307/2280095.

[143] Brier GW. Verification of forecasts expressed in terms of probability. Mon Weather Rev 1950;78:1–3. http://dx.doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

[144] Hosmer DW, Lemeshow S. Goodness of fit tests for the multiple logistic regression model. Comm Statist Theory Methods 1980;9:1043–69. http://dx.doi.org/10.1080/03610928008827941.

[145] Demšar J. Statistical comparisons of classifiers over multiple data sets. J Mach Learn Res (JMLR) 2006;7:1–30.

[146] Kruskal WH, Wallis WA. Use of ranks in one-criterion variance analysis. Am Stat Assoc 1952;47(260):583–621. http://dx.doi.org/10.1080/01621459.1952.10483441.

[147] Drummond M, et al. Methods for the economic evaluation of health care programmes. Oxford University Press; 1997.

[148] Nemenyi PB. Distribution-free multiple comparisons. ProQuest Diss Theses 1963;127.

[149] Keselman HJ, Rogan JC. The Tukey multiple comparison test: 1953–1976. Psychol Bull 1977;84:1050–6. http://dx.doi.org/10.1037/0033-2909.84.5.1050.

[150] Alabdulhafith M, et al. A clinical decision support system for edge/cloud ICU readmission model based on particle swarm optimization, ensemble machine learning, and explainable artificial intelligence. IEEE Access 2023;11:100604–21. http://dx.doi.org/10.1109/ACCESS.2023.3312343.

[151] Krumholz HM, et al. Relationship between hospital readmission and mortality rates for patients hospitalized with acute myocardial infarction, heart failure, or pneumonia. JAMA 2013;309:587–93. http://dx.doi.org/10.1001/JAMA.2013.333.

[152] Moon TK. The expectation-maximization algorithm. IEEE Signal Process Mag 1996;13:47–60. http://dx.doi.org/10.1109/79.543975.

[153] Radford A, et al. Language models are unsupervised multitask learners. 2019.

[154] Sánchez-Maroño N, Alonso-Betanzos A, Tombilla-Sanromán M. Filter methods for feature selection – a comparative study. Lect Notes Comput Sci (Incl Subser Lect Notes Artif Intell Lect Notes Bioinform) 2007;4881 LNCS:178–87. http://dx.doi.org/10.1007/978-3-540-77226-2_19.

[155] Kohavi R, John GH. Wrappers for feature subset selection. Artificial Intelligence 1997;97:273–324. http://dx.doi.org/10.1016/S0004-3702(97)00043-X.

[156] Guyon I, Elisseeff A. An introduction to variable and feature selection. J Mach Learn Res 2003. http://dx.doi.org/10.5555/944919.944968.

[157] Lal TN, Chapelle O, Western J, Elisseeff A. Embedded methods. Stud Fuzziness Soft Comput 2006;207:137–65. http://dx.doi.org/10.1007/978-3-540-35488-8_6.

[158] Wang S, et al. MIMIC-extract: A data extraction, preprocessing, and representation pipeline for MIMIC-III. In: The ACM conference on health, inference, and learning. CHIL '20, New York, NY, USA: Association for Computing Machinery; 2020, p. 222–35. http://dx.doi.org/10.1145/3368555.3384469.

[159] Tang S, et al. Democratizing EHR analyses with FIDDLE: a flexible data-driven preprocessing pipeline for structured clinical data. Am Med Inform Assoc 2020;27(12):1921–34. http://dx.doi.org/10.1093/jamia/ocaa139.