Length of stay prediction using machine learning for Intelligent and efficient hospital management and the associated challenges: A systematic review

Lyse Naomi Wamba Momo *[a], Giulia Rinaldi[a], Xi Shi[a], Frank Rademakers[b], and Bart De Moor[a]

[a]KU Leuven, Department of Electrical Engineering (ESAT), Stadius Center for Dynamical Systems, Signal Processing and Data Analytics, Leuven, Belgium
[b]Division of Cardiovascular Imaging and Dynamics, Department of Cardiovascular Sciences, KU Leuven, Leuven, Belgium

February 15, 2022

*Corresponding author: lysenaomi.wambamomo@kuleuven.be
Address: ESAT-STADIUS, Stadius Centre for Dynamical Systems, Signal Processing and Data Analytics, Kasteelpark Arenberg 10, 3001 Leuven, Belgium.

Abstract

Motivation: Inpatient length of stay (LoS) has proven to be an efficient management tool which if monitored continuously, can help hospital management systems to improve resource allocation, maintain optimal patient flow and increase patient satisfaction. Its relevance has attracted a lot of attention from a number of authors trying to construct and validate prediction models capable of capturing the real-time dynamic of inpatients.

Objectives: The First objective of this work is to report on the most recent advances towards LoS prediction focusing on six aspects; formulation of the prediction problem, data structures used in modelling, solutions to complex challenges arising when modelling patient data, models developed or used, model evaluation and interpretability. Secondly, we also assess the approach used in the selected studies towards improved hospital management.

Methods: From 2411 publication initially retrieved in six digital libraries, 38 primary studies were retained after applying inclusion and exclusion criteria.

Results: The analysis of these primary studies showed that while most authors used the classification approach for LoS prediction, binary or 3-classes classification are not adapted for efficient hospital planning while increasing risks of class imbalance that can greatly mislead model accuracy. The most used model was random forest with reported good performance. Linear and logistic regression were mostly used as baseline models and in most cases the former had the worst performance. Sequence-to-sequence neural networks that have the ability to handle the time dimension reported the best performance for the continuous prediction of LoS and were the most used deep learning architectures. One of the foremost contribution of the reviewed studies was the use of attention mechanisms with sequential architectures or solely like in transformer networks that provide intepretability of powerful black-box models. This review also showed that the use of multi-modal data structures greatly improved model accuracy while taking into account the correlations between patients using graph neural networks. The most used metrics were the Area under the receiver operating curve (AUROC) and mean absolute error (MAE) for classification and regression tasks respectively. The majority of researchers reported on the most predictive factors towards LoS prediction with feature importance from random forest or XGBoost being mostly employed. Truncation of the outcome distribution was mostly used to reduce the effects of skewness and imputation most done by values.

Conclusions: The main contributions of the primary studies included the development and adaption of novel architectures for LoS prediction and the adaptation of the outcome to increase utility by management teams. However, some limitations were also observed in terms of the prediction approach, the choice of the clinical outcome, the patient populations and the lack of model interpretability especially for black-box models. Finally, eight suggestions are made to improve the research towards LoS prediction.

1 Introduction

The demand for healthcare is gradually on the rise worldwide [1]. The healthcare sector will occupy more than 20% of the Gross Domestic Product (GDP) of several countries by 2050 due to among other reasons, the sub-optimal use of healthcare resources [1, 2]. Sudden events such as the COVID-19 pandemics have largely exacerbated this demand [3] during the past two years. In some EU countries (EU-27), the greying phenomenon also contributes to the expenditure increase [4] due to the increase in concomitant chronic diseases with age, while causing a reduction in the healthcare workforce [5]. A study carried out by Swensen, Strongwater, and Mohta [6] reveals that about 82% of burn-out among this healthcare workforce is directly associated to the health system organization. Thus, the constant provision of healthcare services when there is a shortage in workforce and increase in demand could be improved by more efficiently planning hospital activities.

Efficient planning can be achieved by monitoring indicators of effective healthcare delivery like inpatient length of stay (LoS). This indicator is often used to compare the level of care coordination across countries

within the same geographical area and having comparable levels of technology [7]. LoS, defined as the duration of inpatient stay from admission to discharge in a single care episode, is a powerful tool to assess the efficiency and effectiveness of hospital management [8]. Because hospitals have as core mission the provision of accessible and high quality care within cost containment, it is required that all healthcare services are efficiently planned, from admission up to discharge. That is, after diagnostic workup, surgery or other forms of treatment are administered, rooms must be allocated, available staff must be adequately managed and resources allocated in proportion to the needs of the patient. Measures like LoS can help measure hospital throughput when all these components are effectively handled together. Real-time and accurate LoS estimations, that is, bed occupancy prediction can help maintain an optimal patient flow [9, 10] given that hospital capacity is fixed and known. A margin for unplanned events, defined using statistical measures like the mean, median or more robust measures of the frequency of their past occurrences, can even be included in the planning. This will provide the hospital with the adequate resources in such circumstances.

The importance of LoS as a measure of resource consumption and hospital (or medical unit) performance has prompted clinicians to constantly predict hospital LoS [10] or intensive care unit (ICU) LoS [11] for timely hospital management. In the latter, physicians were only able to correctly predict short stay patients of less than five days with an overall accuracy of only 52.7%. As reported by this study, the manual prediction of LoS reached even lower accuracy for medical and emergency surgery patients [11] and those with longer expected ICU LoS. Vicente et al. [12] suggested that prediction accuracy increases with seniority in the medical field. The increasing use of artificial intelligence (AI) in the domain of healthcare indicates possibilities of improved prediction of LoS. Barnes et al.[9] found that predictive models for LoS achieved higher prediction accuracy for near-future discharge predictions compared to clinicians predictions and used in less time [9].

Input data used for LoS modelling and prediction are often stored in voluminous databases called electronic health records (EHRs) [13]. EHRs are a digitized version of patient files enriched on a daily basis that have the potential to help medical institutions make informed and tailored decisions about a patient's care [13]. Recently, several studies have employed a myriad of tools from traditional machine learning algorithms to complex and state-of-the-art deep learning architectures for building LoS predictive models, aiming to provide means for more efficient hospital resource allocations [14–16], improvement in patient care and satisfaction [17, 18], stratification and early identification of patients at risk of extended stays [19, 20]. Sometimes, these complex models have been stacked or assembled with each other in order to strive for high prediction accuracy and more importantly, handle different challenges encountered when modelling EHRs. These challenges include right skewness in the outcome of interest, sparsity and temporality in the data, irregular measurements of patient parameters, missing information, heterogeneity in patient populations, erroneous, inconsistent and outlying observations, interpretability of complex models and real-time predictions. The latter being of utmost importance for hospital management as it permits to detect and adjust plannings when unforeseen circumstances that may extend stays occur.

In contrast to the considerable number of research papers found under LoS prediction models, only one survey [21] and two review papers [22, 23] were encountered that summarize work done under the topic of interest. While the survey paper focuses on prediction model types by partitioning existing work into the broader analytical approaches used, the review from Bacchi et al.[22], gives an overview of different research papers with their limitations and Atashi et al.[23] elaborates on different aspects involved in the development of linear ICU LoS predictive models for patients who underwent coronary artery bypass grafting. None of these papers discusses the underlying mechanisms of the different models used, the influence of the different approaches used on overall model performance nor the associated challenges encountered during modelling. Also, given the increasing attention of healthcare as an area of application of novel AI architectures, these surveys turn to be outdated. This review therefore, aims to address the afore-mentioned limitations by looking at both traditional and state-of-the art machine learning models used on publicly available data

and discusses both their underlying mechanisms as well as their suitability for real-time LoS predictions, including proposed solutions to identified modelling challenges. Moreover, it assesses the feasibility of a real-life implementation of these developed models for the purpose of hospital management. That is, for the purpose of planning all day-to-day activities several days ahead and not as a discharge decision meter in lieu of medical unit heads. For example, predicting patient stays as being short or long if LoS < 7 or ≥ 7 days respectively as the great majority of authors did, does not give adequate information to the management as to the planning or admission of new patients, planning of surgeries, ordering of additional resources, checkups to be done or not, a critic raised by other authors [14, 15, 24, 25].

By focusing on work done on publicly accessible data in which authors set out to apply and compare existing models or develop new robust models against inherent challenges in the data, this study therefore presents a systematic literature review that summarizes and critically analyzes recent work on LoS prediction models under the following aspects; approach to the LoS problem, structure of the data used for modelling, investigated remedies to common challenges encountered in modelling EHRs, models used or developed, performance evaluation metrics and achieved performance, model interpretation and validation. Moreover, the accessibility to the used data offers the possibility of replicating the approach in presented studies.

Section 2 details how the systematic review guidelines were implemented to narrow the set of papers on LoS prediction to those that specifically answer the research questions. Section 3 discusses all main findings extracted from published papers. In section 4, the main contributions and limitations of the reviewed studies with respect to hospital management are discussed. Some research perspectives are proposed in section 5 and section 6 concludes the work.

2 Research Methodology

This systematic literature review was conducted in accordance with the guidelines outlined in Keele et al. [26]. These guidelines were used to define a search strategy that enabled to trace out recent studies that make use of machine learning and deep learning tools for LoS prediction models. This strategy included; paper search and screening, information extraction and information synthesis. The search and screening phase involved retrieval of papers from journal archives using various combinations of search keywords and the application of inclusion and exclusion criteria. The information extraction phase involved careful and organized extraction of data from primary studies which were then synthesized for reporting in a comprehensible way and critically analyzed in the synthesis phase.

2.1 Search method

A systematic paper search was carried on independently by two authors on six digital libraries: Scopus, Web of Science, PubMed, IEEE Xplore, Google Scholar and Science Direct, using a combination of search terms defined as (LoS "OR" Management strategy) "AND" Prediction model. The specific terms used are outlined in Table 1. The search was limited to studies published between January 2016 and April 2021.

Table 1: Research Keywords

Category	**Search term**
LoS	Length of stay, inpatient stay, duration of stay, patients length of stay
Prediction Models	Length of stay annotation, length of stay prediction, length of stay classification, clinical outcome prediction, length of hospital stay prediction, length of stay prediction using clinical notes, length of stay deep learning
Management Strategy	Hospital management

2.2 Inclusion and exclusion criteria

After removing duplicates from the papers downloaded, inclusion and exclusion criteria were applied as first filter to eliminate irrelevant papers.

2.2.1 Inclusion criteria

A published paper is included if;

1. The manuscript is published between January 2016 and April 2021.
2. The paper is a journal, conference, proceeding or book section.
3. Machine learning and/or deep learning algorithms are employed for LoS predictions.
4. Publicly available, open-access or authorized data is used.
5. Inpatient LoS is either the sole outcome or among other outcomes, e.g. [14] or as part of a composite outcome in a multi-task framework, e.g. [14, 27–29].

2.2.2 Exclusion criteria

A paper is excluded if;

1. It only statistically compares the difference between observed and physicians predicted LoS, e.g. [30]
2. The manuscript only discusses or analyzes factors affecting LoS without constructing a prediction model for estimating new patients LoS in a test set, e.g. [31–33]
3. The outcome of interest is not inpatient LoS, e.g. [34].
4. LoS is used as a covariate and not the outcome variable, e.g. [35]
5. Only linear machine learning models are employed, that is, not as baseline but as sole models. Most often, when applying only linear regression, e.g.[36, 37] or logistic regression, e.g. [38–40], the objective is to study the factors affecting LoS and also, these linear models cannot adequately capture the complex interactions within the patient data.
6. Baseline characteristics of patients included in the study are not mentioned, e.g. [41]
7. Results reporting is unclear, e.g. [42, 43]
8. Access to the data used in the study is subjected to a fee, e.g. [19, 44, 45].

2.3 Data extraction and synthesis

In total, 2,411 papers were extracted from six digital libraries and 1,022 duplicates removed from these. Exclusion criteria were applied in three filtering steps one after the other. Abstract screening in the first filtering step resulted in the exclusion of 1,073 papers. Second filtering of 316 papers in full text to identify the nature of the predictive model excluded 159 others. Finally, out of 157, only 36 primary studies used publicly available, open-access or authorized datasets. Eleven authors were contacted as the nature of the data used was not clear but only three replied as working on private datasets, so studies from eight other authors were not included. Two other studies were included after consulting the reference list of thirty-six previously included studies. The search and extraction processes are summarized on Figure 1.

Excel sheets were created and populated with information extracted from all thirty-eight selected primary studies. This information is disseminated in graphical, tabular and narrative forms in Section 3.

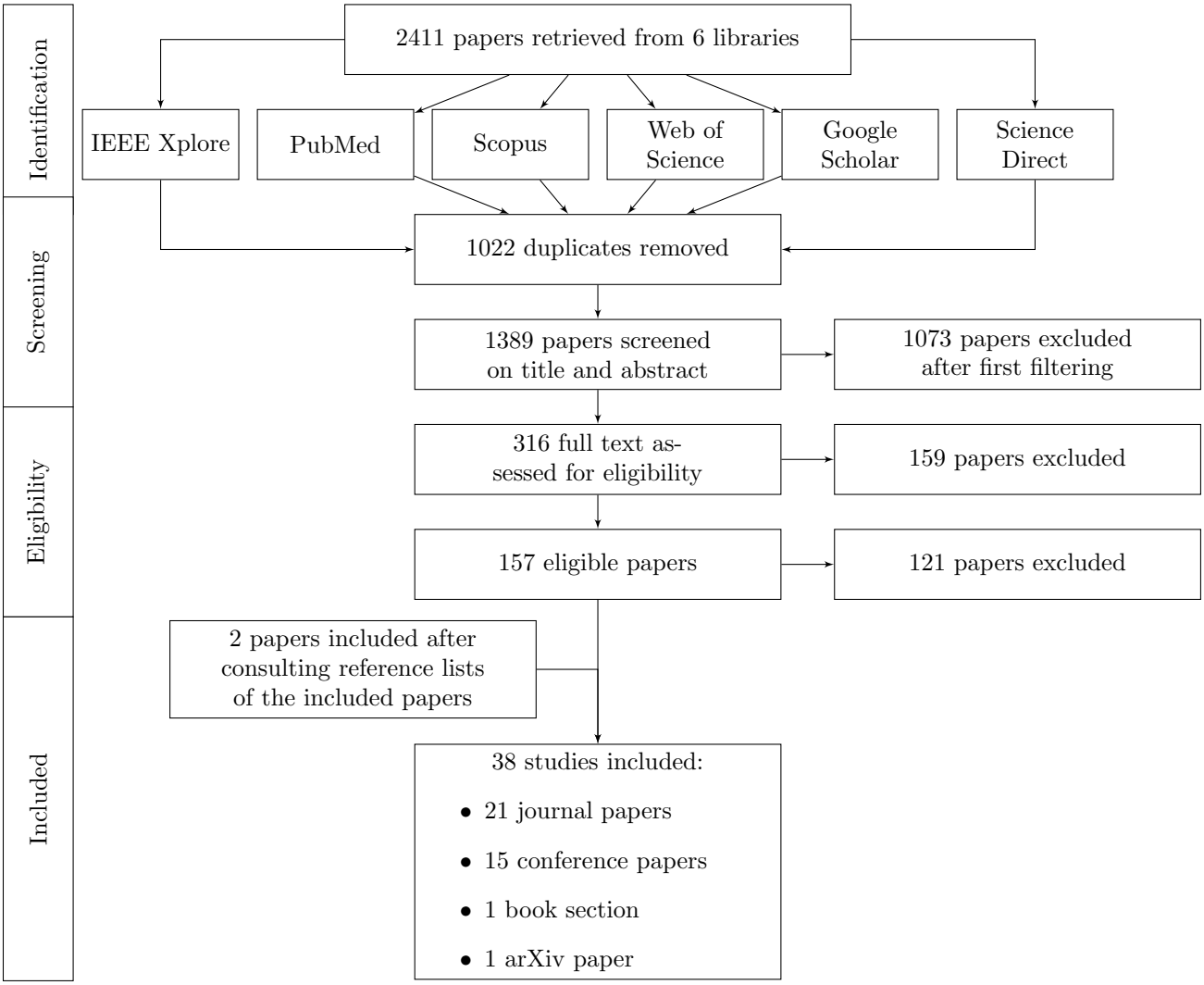


Figure 1: Flow diagram for studies identification, screening, eligibility and inclusion.

3 Results

The information retrieved from published papers is summarized and analyzed in this section. This review aims to simplify researchers work interested in different machine learning and deep learning prediction models used so far for LoS by synthesizing information from previous work, providing a precise and critical idea on different approaches towards hospital management via LoS predictions.

Though many studies considered multiple clinical outcomes like risk of mortality, physiologic decompensation (risk of rapid deteriorating health condition in the next 24 hours), phenotype classification, this study only reports on the LoS prediction outcome due to its complex nature.

3.1 Main findings

Thirty-eight studies, among which twenty-one journal papers. fifteen conference papers (C), one book section (BS) and one arXiv papers (see, Figure 2) were scrutinized.

LoS is defined as the length of an inpatient episode of care, calculated from the day of admission to the day of discharge and based on the number of nights spent in the hospital [8] or the total number of days measured in multiples of 24 hours that a patient occupies a hospital bed [46]. Per these definitions, LoS can be calculated on a hospital level as well as on a unit level like ICU where demand and early provision of critical

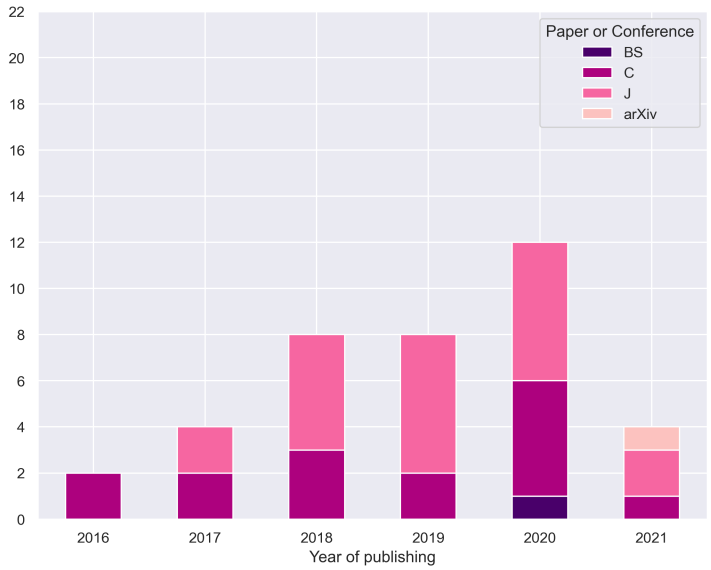


Figure 2: Publication Trend up to April 2021

care is crucial for patients' survival. Twenty seven out of thirty-eight studies used ICU data for predicting either ICU LoS or hospital LoS. Four out of these twenty seven focused on particular ICU sub-populations, neonatal ICU [47], surgical ICU [48], ICU patients with heart failure [49] and ICU patients diagnoses with sepsis [50]. One study use Covid-19 patients data [51] and only ten out of thirty-eight use acute care ward data. Regarding the outcome, nineteen predicted hospital LoS using acute care ward data [17, 52–55] or ICU data [18, 27, 49, 51, 56–65] and fifteen predicted ICU LoS. One study, Balkan and Subbian[59] predicted both hospital and ICU LoS, and another, Belderrar and Hazzab[66] predicted hospital high LoS outliers defined using the geometric mean plus two standard deviations. Finally, two studies did not explicitly mention the area of prediction [67, 68].

Seven out of thirty-eight studies used data collected across multiple centres or hospitals [14, 61, 62, 68–71]. MIMIC-III database [72] was the mostly used unicentric dataset appearing in eighteen studies. Uniform resource locators (URL) of each of the database used, access conditions as well as currently available extraction pipelines are detailed in Tables 2 and 8 respectively. Varying population sizes were used with a minimum of 477 newborns undergoing the Noorwood procedure [54] and a maximum of 378 016 patients from 125 medical units and 14 hospitals in Wang et al. [62].

As shown in Figure 4, the majority of the studies, 73%, used demographic patient information as one of the input features to the predictive models. This includes, age, gender, ethnicity, insurance type, etc. Diagnoses codes among other features were used in 28% of the studies [14, 17, 47, 49, 50, 56, 59, 61, 65, 69, 71]. While some authors used very few patient features [63, 67, 68], others used a plethora of them from diagnoses, demographic, laboratory measurements, vital signs and clinical notes. Three out of thirty-eight authors used data of different modalities by combining clinical notes to demographic and time-varying features to improve prediction accuracy, [18, 29, 73]. One study used 1399 binary drug features as sole set of predictive variables [70].

Though this outcome is measured on a continuous scale, the prediction problem can be re-formulated depending on the research objectives. Of the 38 consulted studies, 23 reformulated the prediction problem into a classification problem, either binary or multi-class by splitting the outcome into buckets and predicting the probability of a patient stay falling in one of these. Twelve used the outcome in its original state by predicting the number of days till discharge in a regression framework. Two studies used a two-step approach where first, a classification or clustering algorithm was applied to isolate a group of patients and then performed a

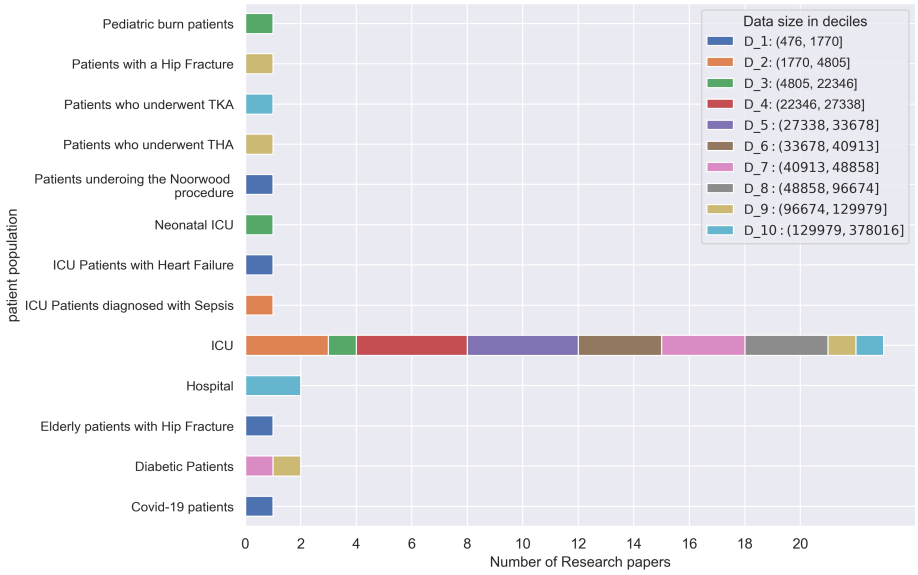


Figure 3: Number of studies with different patient populations investigated and associated patient sizes (divided in deciles). Surgical ICU population from Batista and Sanchez-Arias[48] doesnot appear on Figure because data size was not reported. TKA: Total Knee Arthroplasty, THA: Total Hip Arthroplasty, ICU: Intensive Care Unit

regression on this isolated group [62, 74]. One study predicted the probability of discharge over time (time to discharge) using machine learning approaches to survival analysis [51]. Among the 23 studies that considered the classification scheme to the LoS prediction problem, 11 performed binary classification predicting patient stays as either short or long stays using different threshold values, with seven days being the most used threshold to delineate stays [18, 28, 29, 49, 57]. One-third of the studies that used the multi-classification approach split the outcome in ten classes, stays of less than a day, stays of one to seven days, stays longer than a week but less than two weeks and finally patients that spent more than two weeks within ICU [73, 75, 76].

Mostly used predictive models for LoS predictions among other tasks are shown in Figure 5. Random forests (RF), logistic regression (LR) and linear regression (Linear R) were used with other algorithms in 37%, 26% and 21% of the studies respectively. Standard deep neural networks, recurrent neural networks like long short-term memory (LSTM), gated recurrent units (GRU) and their variants were used in respectively 24%, 21% and 13% of reviewed papers. Boosted trees were also widely used either for regression or classification. Gradient boosted trees (GBoost) and extreme gradient boosted trees (XGBoost) were used in 18% and 8% of the studies respectively. Bagging of classification trees (excluding random forests) were used in Purushotham et al.[58]. Stacking of different base learners (algorithms) together was carried out in three studies [17, 49, 50]. One study employed reinforcement learning by training two agents in order to mimic instances where patients seek second opinions from other medical practitioners [77].

Regarding model evaluation using performance measures in the classification framework, the area under the precision-recall curve (AUROC) was used in 23% of the studies, the accuracy measure in 15%, followed by the F1-score, Cohen's linear weighted kappa, precision and recall in 10%, 8%, 6% and 6% respectively. In the regression approach, 24% of the studies used the coefficient of determination R^2 and 18%, the mean absolute error (MAE). The mean squared error (MSE) and the root mean squared error (RMSE) were used in 12% and 9% of the studies respectively. Another means of model evaluation, which is benchmarking of developed and novel architectures on a new dataset to access the robustness of the model was carried out in Rocheteau, Liò, and Hyland [14] and Che et al.[27]. The reported performance by all authors using the

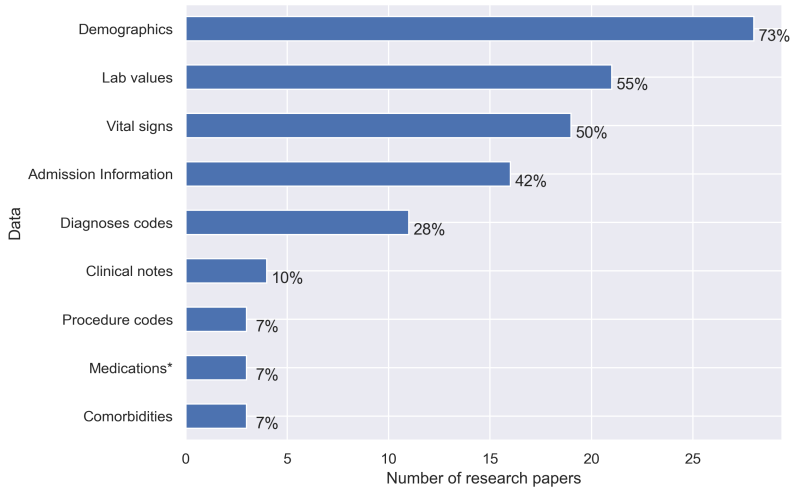


Figure 4: Categories of features used in more than two studies. (*) Denotes count of feature.

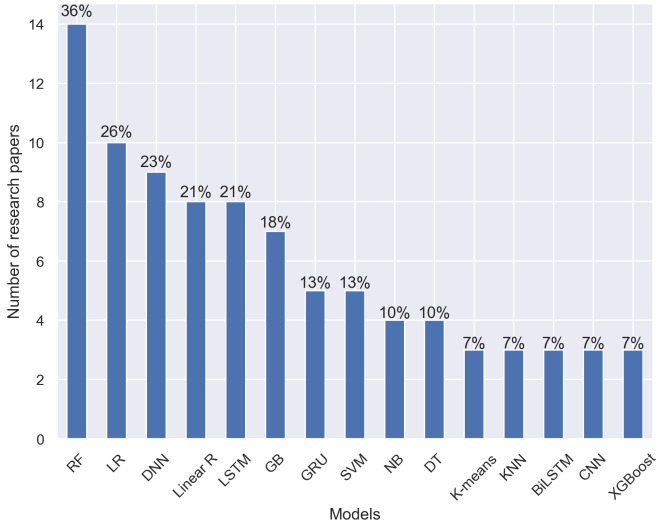


Figure 5: Models employed in more than three studies

various metrics are given in Table 9.

As a means of model interpretability, twenty out of thirty-eight studies either reported on the most influential features towards LoS prediction or unveiled the black-box nature of the algorithm used so that factors driving predictions for a particular patient were showned.

3.2 Data - Electronic Health Records (EHRs)

As seen in Figure 4, EHRs typically contain all possible data fields describing the patient health trajectory over their entire stay. From historical information, demographics and admission data, vital signs, laboratory results, nurse assessment reports, prescribed and administered medications, procedures to transfers, costs and discharge destinations. Popular publicly available EHRs like MIMIC-IV [78] or eICU-CRD [79] contain longitudinal patient data recorded over more than a hundred features. In the case of MIMIC-IV, additional tables for emergency department [80] and chest labeled radiographs [81] are also available that can be linked to MIMIC-IV using unique patient identifiers. This then leaves the researchers with a myriad of features, all potentially useful or not for clinical outcome modelling and prediction. However, extracting, curating and pre-processing all these different data types (static and time-varying numerical and categorical inputs, text

and images) for all patients so they can be fed into a model is quite a tedious task. Whether all these data sources are indeed relevant for modelling or how to select relevant features for the specific outcome then stand as additional research questions.

In the consulted studies, a great number of authors chose input features following previous studies, re-using already available data extraction and pre-processing publicly released pipelines made available in Table 8. The drawback of this approach is that it tends to generalize the features that could impact LoS where as, different parameters are usually monitored on different patients following their conditions. In section 4.2.3 the limitations of the apriori choice of model input features are discussed and alternatives are proposed in section 5. The various datasets used in the consulted studies, their url, and description are summarized in Table 2.

3.3 Inherent Challenges when dealing with EHR and Investigated Solutions

When extracting patients′ information from EHRs, this usually comes as sparse longitudinal data where for each patient and at specific time points, different hospital parameters are recorded with their corresponding values. This often results in irregularly sampled time-series within and across patients which intensifies when all data are combined. This further increases when different data modalities (numeric, images, text, etc.) are used, the number of patients increase and the patient population is more heterogeneous.

3.3.1 Skewness in the outcome

The distribution of inpatient LoS is always highly skewed to the right with the majority of patients experiencing short stays and very few experiencing long stays. The skewness of this distribution poses a problem when modelling since longer stays (outliers or extremities in data) are often difficult to learn by the model especially when early available data, e.g., first 24 or 48 hours are extracted for modelling. Also, when the patient LoS prediction has to be recalculated during the hospital stay like when predicting remaining time in ICU as done in [14, 69, 73, 75–77] , this minority of extreme stays increases computational complexity since the dataset dimension increases with time. Solutions proposed by various researchers to mitigate the effect of skweness on overall model performance are summarized in Table 3.

3.3.2 Imputation

Missing information is often encountered in data related studies including EHRs for multiple reasons [84]. After extracting data from EHRs, a unified pattern must be adopted for all patients, all data modalities and for each time point such that a well structured tabular dataset can be fed into the predictive model. Re-sampling this data to a common time base (e.g. hourly or daily) so that there is regular spacing between measurements further increases the missingness rate as input parameters are not all measured at the same time point (e.g., within the same hour of the same day). Table 4 gives a summary of different imputation strategies used by various researchers. An extensive study and comparison of existing and novel imputation strategies for recurrent neural networks modelling of multivariate time series is performed in Che et al.[27].

3.3.3 Multi-modal modelling and Irregular Sampled data

Patients are monitored over multiple hospital parameters during a stay. Some of these parameters can be recorded continuously using high-frequency monitors, (e.g., oxygen saturation, heart rate, blood pressure, etc.), some charted at irregular time intervals by nurses or doctors, (e.g., notes on patient clinical assessment, radiology reports, etc.), some that do not change during the patient stay (e.g., age, gender, etc.) and others recorded only once or twice in a day (e.g., laboratory tests). This then results in varying information density across different data structures. An example mentioned in Xu et al. [82] is that an ICU patient can

Table 2: Publicly accessible datasets used by various researchers

Dataset	URL	Description	Studies
MIMIC-II (Multi-parameter Intelligent Monitoring in Intensive Care)	https://archive.physionet.org/mimic2/	Open available database for inpatient physiological signals and vital signs time series captured from patient monitors for over 23,000 adult ICU patients. Data were collected between 2001 and 2008 from a variety of ICUs in a single tertiary teaching hospital.	[74]
MIMIC-III (Medical Information Mart for Intensive Care III)	https://mimic.mit.edu/docs/iii/tables/	Openly available unicentric dataset of over 60,000 ICU admissions for 53,432 adult and 8100 neonatal patients recorded between June 2001 and October 2012.	[18, 28, 29, 47–50, 56–58, 65, 66, 73, 75–77, 82, 83]
MIMIC-IV	https://mimic-iv.mit.edu/docs/access/	Update of MIMIC-III with data of over 200,000 patients admitted to the emergency department and over 60,000 ICU stays in a single tertiary hospital between 2008 and 2019.	[14]
MIMIC-III Waveform database	https://physionet.org/content/mimic3wdb-matched/1.0/	freely-available dataset of 67,830 records for over 30000 ICU patients with most record sets including digitized signals of ECG, ABP, HR, etc.	[82]
eICU (Philips eICU Collaborative Research Database)	https://eicu-crd.mit.edu/about/eicu/	Multi-center clinical database with over 200000 admissions for over 139,000 unique patients admitted to 335 units at 208 hospitals in the USA between 2014 and 2015.	[14, 59, 69, 71]
Physionet 2012 Challenge	https://physionet.org/content/challenge-2012/1.0.0/	Publicly available data for ICU mortality prediction challenge for 12000 described by 7 static and 37 time-series variables. Only first 48 hours in ICU were made available.	[27, 55, 60]
Heritage Health prize data	https://foreverdata.org/1015/index.html	Publicly available 48 month period claims data of over 147,000 records on 139 features	[52]
UCI diabetes dataset	https://archive.ics.uci.edu/ml/datasets/diabetes	Data of 101,766 diabetic patients over 50 features obtained from an automatic electronic recording device and paper records at 130 US hospitals between 1999 and 2008.	[17, 61]
Irish Hip Fracture data	https://www.noca.ie/irish-hip-fracture-database	Clinically led, web-based audit which measures the care for all people over the age of 60 admitted to hospital with hip fractures, across the 16 operating hospitals.	[64]
NewYork State Department of Health's Statewide Planning and Research Cooperative System (SPARCS)	https://www.health.ny.gov/statistics/sparcs/#documentation	Data OF 103,592 patients from the New York State Department of Health's SPARCS database, collected on patient-level and all discharges from hospitals and clinics in New York between 2012 and 2016	[53, 67, 68]
Pediatric Heart Network (PHN) that conducted the Single Ventricle Reconstruction (SVR) Trial	https://www.pediatricheartnetwork.org/ForResearchers/PHNPublicUseDatasets/SingleVentricleReconstruction-SVR	Data of 920 newborns with single ventricle physiology collected from 15 centers between 2005 and 2009 by the PHN SVR Trial investigators.	[54]
Web-based international burn registry of German- speaking countries	https://verbrennungsmedizin.de/verbrennungsregister/	Data of 8,915 pediatric burn patients obtained from the Web-based international burn registry of German- speaking countries (Germany, Switzerland, and Austria) from the German Society for Burn Treatment (DGV) between 2015 and 2018	[63]
Program for medicalization of information systems in medicine, surgery, obstetrics and odontology	https://epidemiologie-france.aviesan.fr/en/epidemiology/records/programme-de-medicalisation-des-systemes-d-information-pour-les-soins-de-courte-duree-en-medecine-chirurgie-obstetrique-et-odontologie	French national information system for 378,016 inpatient stays registered at Hospices Civils de Lyon (HCL) in 125 medical units from 14 hospitals of the HCL in 2013	[62]
Open-access data	https://www.thelancet.com/journals/laninf/article/PIIS1473-3099(20)30119-5/fulltext	Data from 1,182 Covid-19 patients	[51]

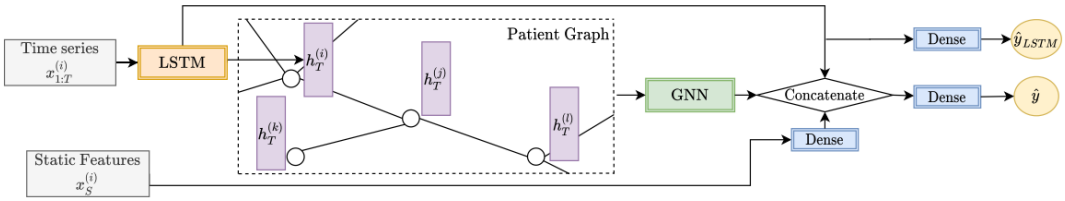
Table 3: Possible remedies to skewness in LoS proposed by various authors

Solution	Description	Studies
Log-transform	for centering the distribution so that it appears symmetric as a *normal distribution*. Commonly used with traditional machine learning models	[49, 61]
Loss function	choosing a loss function like mean-squared logarithmic error (MSLE) for training models. MSLE is more adapted for capturing extremities in right skewed data than the mean absolute error (MAE)	[14, 71]
Truncation	Truncating the distribution at a particular value to reduce the effect of extended stays on overall model accuracy.	[62, 64, 65, 77]
Categorization of the outcome	conversion of the original outcome to a categorical one using a threshold defined using a robust measure like the median or using multiple thresholds	[17, 55]

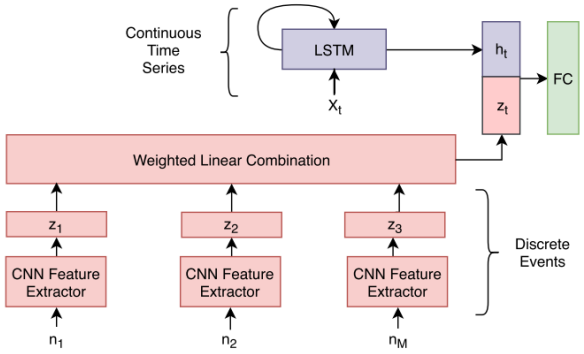
Table 4: Imputation strategies proposed by various authors

Solution	Description	Studies
By Values	previous recordings for the same patient within the same day are copied and forwarded or backwarded by assuming they did not change too much within a short time. Imputation using known standard values for *healthy* patients, using the mean or median for numeric features and the mode for categorical features, using a new value like 'NA', imputation with zero after normalization, using median values	[14, 18, 28, 50, 55, 58–60, 69, 76, 82]
Mask Vector	binary vectors created and attached to each input variable meant to indicate at each time step if the given entry is an imputed one or not. That is, the input dimension is multiplied by two	[76, 82]
Decay Indicators	similar to mask vectors but additionally contains a decay value that indicates the time since the last non-imputed recording	[14, 27, 28]
Multiple Imputation	missing data approximated using a range of values instead of one value using Markov Chain-Monte Carlo (MCMC), from previous studies	[49, 54]
Automatic imputation using available packages	*missforest package, Amelia software package* for missing data imputation	[17]
Asexual reproduction optimization (ARO) algorithm	combination of genetic algorithm and ARO	[47]

generate up to 11 million values from a single lead ECG recording sampled at 125HZ and 86 thousand vital signs sampled minutely within a day. Because LoS prediction requires modelling all these data modalities, with different structures and frequency of measurement among and across patients, together, representations from highly dense data could hide valuable information from sparse data, [82]. Nevertheless, a number of authors, [18, 28, 65, 73, 82] have shown that using multiple data modalities can improve model accuracy. The approach used here involved the use of separate models for learning the different data structures which representations were later concatenated to make predictions. To cite a few, Xu et al. [82] used separate architectures to process different patient parameters, later combined and fed to a multivariate regression model for predictions. While high-density waveforms from patient ECG and vital signs were embedded into a low and rich representation using a channel-wise CNN and then passed through an LSTM to encode temporality in the data, summary statistics (e.g. mean, maximum, minimum, etc.) were extracted from hourly charted records. Static demographic variables were left untouched and onsets of laboratory tests and interventions (like medication administration) were included in binary matrices with time indexing indicating



(a) LSTM-GNN model: LSTM representations from time series features $x^{(i)}_{1:T}$ concatenated to GNN embeddings of diagnoses and static features $x^{(i)}_S$ embeddings from fully connected layer.



(b) 1D CNN-LSTM model: z_i are CNN representations from clinical notes n_i and h_t, LSTM representations from time series x_t

Figure 6: Multi-modal model architectures from (a) Rocheteau et al.[71] and (b) Khadanga et al.[73]

at what time an onset occurred, and then used for guidance via attention mechanisms. In Khadanga et al. [73], hourly charted time varying data (17 time-series data used in Harutyunyan et al. [76]) as well as irregular charted clinical notes were used for building LoS prediction models among other prediction outcomes. At each time step t, clinical notes before time step t were encoded by a CNN and then weighted such that most recent notes are given higher weights in order to capture actual patient state. LoS predictions at time t were then obtained using softmax regression of the average over all notes taken before time t concatenated to the LSTM representations of time series features before time t. In Zhang et al. [18], two multi-modal neural network structures were used to learn deep patient representations from temporal patient signals and clinical notes separately in order to improve prediction performance. In Rocheteau et al. [71], similarities in patient diagnoses were encoded using graph neural networks, which representations were concatenated to those obtained from time-varying and demographic information representations.

As explained before, the time interval separating recordings of patient parameters is irregular since they are not all charted at the same time. Re-sampling and multi-modal modelling were then used to mitigate this effect where the former enables to re-sample them to a common time frame and the latter to not combine all data modalities together. A thorough review on strategies for handling irregularity in the time domain when using gated recurrent neural networks is proposed by Weerakody et al.[85].

3.3.4 Time-series modelling

Since patient information comes in the form of longitudinal sequences ordered in time that describes the patient evolution during the stay, it is important for researchers to take this time dimension into account. The different strategies employed for incorporating this temporality in the data are summarized in Table 5.

Table 5: Different strategies for incorporating the time dimension in the data

Strategies	Description	Studies
Time dependent variable extraction	Extraction of features like the trend, frequency of measurement of each patient parameter,etc. and fit all extracted features to a traditional machine learning algorithm.	[55]
Time-dependent algorithms	Use of recurrent neural networks like LSTM, GRU, temporal CNNs, BiLSTM, BiGRU, etc. that explicitly model temporality in the data	[14, 18, 28, 58, 69, 73, 75, 76, 82]
Positional encoding, Dense interpolation for encoding order	use sinusoidal functions or other types of functions to capture the relative or absolute position of each time-step in the sequence	[14, 28, 75]

3.3.5 Real-time predictions

Few studies employed post-admission data, that is, data available during the patient stay within hospital premises in order to update LoS prediction models during hospitalization and adjust these predictions. Here, we briefly discuss the different strategies employed by some authors to update LoS prediction models using post-admission data.

By extracting the first 24 or 48 hours of patient data, re-sampled on an hourly basis and fed to a recurrent neural network, LoS predictions were computed and updated at each hour during the 24 or 48 hours of stay in [18, 27–29, 47, 56, 58, 60, 71, 83]. A similar work was performed in [14, 69, 73, 75–77] where, hourly predictions of remaining time in ICU were computed between admission and dead or discharge from ICU. These models are therefore updated hourly by combining data at previous time points to data at current time thereby increasing the training dataset as the number of hours within ICU increases per patient, constituting a hinge in computational demands.

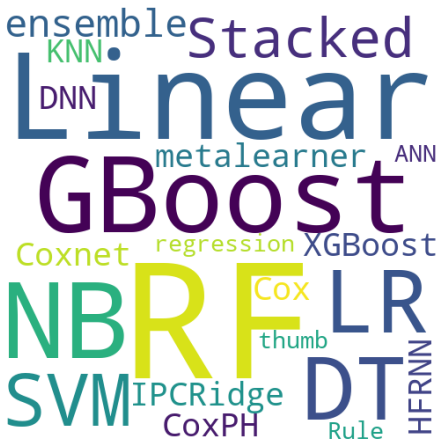
3.3.6 Data density

The most investigated patient population was ICU patients, a rare and expensive resource compared to acute ward data. At least three reasons can account for the choice of ICU data. Firstly, the dense amount of information captured from the continuous monitoring of patients in ICU constitute a good ground for well-calibrated prediction models [86]. Secondly, most publicly available EHRs for benchmarking clinical prediction models comprise mostly ICU patients, [72, 78, 79, 87] and lastly, ICU LoS is also an important management resource for hospital LoS because patients are almost never discharged home following an ICU visit but to an acute ward in which a bed must be made available.

Both the richly sampled ICU data and less sampled data from acute care wards present challenges. Some challenges with ICU data are associated with scability and computing resources since the abundant information can sometimes require huge memory resources when loading and manipulating during pre-processing. Also, the abundant nature of the data makes it suitable for data demanding models like deep learning architectures. Whereas in acute care wards, some challenges are associated with the sparse nature of the data,



(a) All models used with ICU data



(b) All models used with acute care wards data

Figure 7: WordClouds of models applied on (a) ICU data (b) acute care ward data

which could potentially limit the models used as shown on Figure 7, where LSTM, GRU are employed with ICU data and not with acute care war data.

3.4 Prediction models

This subsection briefly describes the most used algorithms, those that achieved higher performance and how these were adapted to overcome challenges mentioned in Section 3.3. These models are either supervised (requiring two sets of attributes - predictors and target) or unsupervised (requiring only the predictors and whose labels are to be determined). In machine learning, the terms attributes or features are often associated to those characteristics describing the population or sample to be dealt with, which in the clinical context, would be for example, blood pressure, heart rate, temperature, age, etc. The patients themselves whose LoS is of interest are referred to as instances or observations. Each instance is then described by a vector of attribute values where each position in the vector corresponds to a unique attribute [88]. One of these attributes, e.g., LoS, is the target attribute to be predicted by estimating the mapping linking it and all remaining attributes.

3.4.1 Linear models

Linear regression: is a popular method in statistical learning for modelling a quantitative outcome [89]. This method assumes that the sets of attributes are linked by a linear mapping and is formally defined as,

$$Y = \boldsymbol{\beta} X + \epsilon,$$

where ϵ is an error term. By fitting a linear regression model, we seek to find the model coefficients $\boldsymbol{\beta}$ that best approximates the linear mapping. One of the mostly used approach is the minimization of the least squares criterion [89]. The underlying assumptions of linearity and normality in the linear regression model makes it unfit for modelling non-linear problems like inpatient LoS. Due to its inability to model time-dependent data, this model was mostly used as baseline with hand-crafted features and always reported to have the worst performance [17, 47, 52, 58, 62–64, 69, 74, 76].

Logistic regression (LR): is another popular statistical learning method for qualitative outcomes, developed by Joseph Berkson [90]. LR is used to predict the probability of belonging to a class given a set of predictors. The mapping to be learned linking the target and predictor attributes is restricted within 0 and 1; probability

interval. One of such mappings is the logistic function that predicts the probability $P(X)$,

$$P(X) = \frac{\exp \beta X}{1 + \exp \beta X}.$$

The model coefficients β are obtained by the method of maximum likelihood [89]. Since probability $P(X)$ is linear in the coefficients when taking the logit transformation, that is, $\log\left(\frac{p(X)}{1-p(X)}\right) = \beta X$, LR is also not adapted for handling non-linear problems like LoS [91]. LR was used as baseline or for comparison with other algorithms in [18, 28, 29, 48, 50, 61, 75]. Other variants of LR included, Ridge LR and LR with l2 regularization used in Jalali et al.[54] and Gao et al. [77] respectively. In Gong et al.[83], L2-regularized LR was used as main algorithm on feature representations of EHRs.

3.4.2 Decision tree learning

Decision tree (CART - Classification and Regression Trees): decision trees (DT) are recursive algorithms used for classification and regression tasks. Starting from a root node (where all observations are in the same space), the predictor space is split recursively in two branches from top to bottom. To perform this recursive binary split, an optimization algorithm is used to find the best predictor and cutoff point (from all possible values of the selected predictor) such that the residual sum of squares or the classification error rate (or Gini index or entropy) in the split region is minimal. After the first split and the two new regions created, the split process is repeated searching again for the best predictor and best cutoff point that minimize further the loss function. This recursive process continues until a stopping criterion is met [92]. Though DT are easy to implement and understand, they can easily overfit since the training data can be split recursively [92]. Moreover, DT are high variance estimators since little changes on the training data can produce significant variations in model accuracy [93]. The performance of DT models were compared to that of other algorithms in [50, 54, 58, 64].

Random Forest (RF): is a bagging ensemble of decision trees in which multiple independent trees are fit on bootstrap samples from the training data, which helps overcome the high variability in DT. At each split, only a random sample of predictors is chosen as split candidates by each tree in the ensemble [92]. RF was the mostly used algorithm employed among other algorithms and reported to achieve high accuracy by [18, 28, 29, 47–50, 52, 54, 57, 59, 61, 63, 64].

Gradient Boosting trees (GBoost): is an ensemble of trees where instead of fitting multiple trees independently like in RF, trees are fitted sequentially on residuals in place of the outcome of previously grown trees to improve prediction accuracy [89]. GBoost can easily overfit if the number of base learners (decision trees) is not carefully chosen and is slower than bagging tree methods such as RF since tree fitting is done sequentially. However, it is easy to understand and interpret [89]. Studies that used GBoost include [29, 47, 49, 52, 54, 64]. Survival analysis version of GBoost was employed and reported the best performance in Nemati, Ansary, and Nemati[51].

Extreme Gradient Boosting trees (XGBoost): is a modified version of gradient tree boosting. XGBoost are very efficient with high prediction accuracy and can avoid overfitting by using a regularized learning objective that penalizes complex models in addition to feature sampling of predictors like in random forest. It can accommodate a wider range of objective functions than DT and also uses a score function for impurity measurement at each split [94]. XGBoost were employed and performance compared to that of other algorithms in [47, 48, 61].

3.4.3 Traditional DL models

Artificial neural networks (ANN): ANN take their inspiration from the human brain and nervous system. An ANN is a block of structures called nodes, interconnected to each other in a hierarchical manner via weighted

edges and non-linear functions [95]. This gives them the ability to handle non-lineartities and complex interactions in the data [58]. Different types of Neural networks were used by various authors. Standard ANNs such as multi-layer perceptrons (MLP) or fully connected network (FCN) or feed forward networks (FFN) that do not explicitly model the time-dimension and rely on hand-crafted engineered features were used in [49, 54, 57, 58, 64, 65, 69, 70, 76, 77]. ANNs reported the best model performance in Alahmar, Mohammed, and Benlamri[17].

3.4.4 Sequence-to-Sequence models (Seq-to-Seq)

Recurrent neural networks (RNNs): RNNs are a type of deep learning algorithms that can handle temporal or sequential data, processing information and predicting output at each time-step [96–98]. RNNs are very popular for natural language processing (NLP) tasks such as sequence to sequence modelling, translation, text generation, etc. They have a memory structure that retains information at time $t-1$ which is then combined with information at time t for prediction. RNNs-based models used involve;

Long Short-Term Memory (LSTM): Because there could be long sequences of data to remember, variants of RNNs like LSTM [99] have been developed to overcome issues of vanishing or exploding gradients encountered by vanilla RNNs [96–98]. Some of these RNNs-based networks were used with in-built attention mechanisms. Attention in RNN serves a means of interpretability where an attempt is made to identify important time points in the time-series data by receiving higher weights [98]. Since EHRs are essentially sparse longitudinal patient data, RNNs based models were mostly employed either as baseline or as main models. One major drawback of RNNs is that they can be very slow since they are not parallelizable, [100]. LSTM based-models including plain LSTM, channel-wise LSTM , both with deep supervision and their multi-task versions were employed for multi-classification of LoS in Harutyunyan et al. [76]. In a similar manner, Rocheteau, Liò, and Hyland [14] employed LSTM, channel-wise LSTM, transformers, as baseline models against which the performance of a novel model was compared. Sheikhalishahi, Balaraman, and Osmani [69] used a Bidirectional-LSTM (BiLSTM) which is a form of LSTM that processes information in adverse directions simultaneously by using both past and future information to predict current output [96]. They showed that this model outperformed basic linear regression and 1-layer ANN for predicting remaining hours in ICU at each hour. Song et al. [75] compared the performance of linear regression, LSTM models with that of a transformer-based network both in single and multi-tasks settings. Another multi-modal model was constructed in Khadanga et al. [73] with LSTM for modelling the time-series data. Xu et al. [82] used vanilla LSTM-based models as baselines and also compared the performance of a CNN-LSTM model to a multi-modal CNN with multi-channel attention. Other variants of LSTM like Time-aware LSTM (T-LSTM) [101] were used as baseline models in Gao et al. [77]. T-LSTM is an adapted LSTM network with the ability to handle irregular time intervals within time series data by incorporating true elapsed time between sequences of inputs. T-LSTM was developed by breaking the cell memory structure of standard LSTM networks into short and long term memories and discounting the former with decaying weights. Other authors who used LSTM based models include [18, 73]

Gated Recurrent Units (GRU): is another variant of RNNs that avoids issues of vanishing or exploding gradients [99, 102]. In order to learn shared representations from two modalities of data (temporal and static), Purushotham et al. [58] applied GRU on time-series data and a feed-forward network (FFN) on static data. Bi-directional GRU models that model both past and future data simultaneously were also investigated. BiGRU with three attention mechanisms developed in Ma et al.[103] were used as baseline in Gao et al. [77]. GRU with delays (GRU-D) introduced in Che et al. [27] was applied in [14, 28, 71] where mask indicators and time decay curves were added to the network states of GRU for learning long-term dependencies and missing patterns in data. GRU and BiGRU with hierarchical attention weights were also used in Harerimana, Kim, and Jang [56] for learning important representations from patient admission information, free text written

diagnoses, diagnoses and procedure codes.

Transformer Networks: they are models built solely with an attention mechanism that help to capture dependencies in the data [104]. This model has shown promising performance for NLP tasks and is increasingly used in healthcare due to its ability to leverage interpretability by weighting different parts of the data to get their importance. In Song et al. [75], a transformer architecture for multiple output prediction including LoS was developed using positional encodings (to encode the sequential nature of the data) and masked multi-headed attention operation (the sequence of data is masked to indicate the length of historical information to use). This architecture was superior in performance and speed in all prediction tasks. A multi-head self attention transformer model was used as baseline model in Rocheteau, Liò, and Hyland [14]. This algorithm from its inception, has proven to be very effective, especially in terms of speed because it does not make use of recurrence and can process multiple time steps at the same time unlike LSTM or GRU [14, 75, 104]. However, in the realm of time series modelling, they still experience some limitations, particularly for feature extraction [14] and memory demands for very long sequences [85].

3.4.5 Convolution neural networks

Convolution neural networks (ConvNets/CNN): inspired by the human vision system, have been and are still very popular for image-driven tasks like image segmentation and object recognition, see [105, 106] for extensive discussion. Because CNNs can handle array-like inputs of one, two and three dimensions [107], they have been exploited for sequential data like signals [108] or time-series or text data [97] reaching state-of-the art performances. These performances arise from the architecture, especially the convolutional and pooling layers where the former are able to learn the internal representation from the raw data by extracting highly discriminative features from the multivariate time-series without domain experts and capture local interactions [82, 107]. The pooling layers then reduce the dimension of the original series [97] resulting in a lower-dimensional and dense representation of the patient time series reducing the risk of overfitting, the number of parameters for training and consequently training time. A number of authors used CNN-based algorithms as baseline models, for multi-modal learning or as pre-processing tools. In Xu et al. [82], CNN were used for feature extraction and dimensionality reduction for embedding densely-sampled electrocardiogram (ECG) waveform signals. Rocheteau, Liò, and Hyland [14] employed time-series adapted CNN based networks such as temporal convolution network (TCN) and pointwise convolutions for extracting temporal trends in the data and for dimensionality reduction respectively. TCN is a type of CNN that is able to incorporate temporality in data. In order to improve model performance by augmenting vital signs data with clinical notes, Khadanga et al. [73] built a multi-model model by applying a separate one-dimensional CNN (1D-Conv) to note events.

3.4.6 Other Models:

Support Vector Machines (SVM), Naive Bayes (NB), K-Nearest Neighbors (KNN) and K-Means were also employed in at least three studies. SVM, which in some circumstances are comparable to neural networks [61] were used among other models in [48, 50, 60, 61] and variants like Fast SVM and Fast kernel SVM in Nemati, Ansary, and Nemati [51]. Naive Bayes, a Bayesian ML algorithm was used as sole model in [53, 67, 68] and reported the worst performance compared to other models in Alsinglawi et al. [50]. Due to its simplicity in understanding, KNN that classifies new observations based on the class of their N-neighbors, was investigated in [47, 50, 61]. SVM and KNN both had an average performance, being neither the best not the worst performing models. In order to improve prediction performance by focusing on more homogeneous entities, K-means clustering algorithm was used to cluster patients before predicting LoS in Rouzbahman,

Jovicic, and Chignell [74] and in Huang et al. [70] to create patient communities when applying federated learning. It was used as benchmark algorithm in Ma et al. [55].

3.5 Model Assessment

Model performance assessment was carried using evaluation metrics. Some authors implemented employed multi-tasks frameworks in which multiple outcomes were learned together as composite outcomes to improve performance. In Table 6 we report on the best performing developed or used algorithms and best baseline in each study clustered by outcome. For classification schemes, AUROC is reported if used in the study, otherwise, accuracy or kappa measures are used. For regression tasks, MAE is used when available, otherwise, the mean magnitude relative error (MMRE) or R^2 are reported. The concordance index (C-index) was used in the survival analysis approach. The overall model performance on each used model as reported by each author with input variables used are presented in Table 9 in the Appendix section.

Table 6: Best model performances clustered by outcome.
HLoS: Hospital LoS, *ihm*: in-hospital mortality, *ph*: phenotyping, *dc*: decompensation, *cc*: patient had cardiac condition, *surg rec*: patient recovered from surgery or not
FL: Federated Learning, CBFL: Community-based FL, C: channel-wise LSTM, DS: Deep Supervision, MMDL: Multimodal Deep Learning Model, Super Learner-II: ensemble of ML algorithms, GAT: Graph Attention Networks, MPNN: Message Passing Neural Networks, MMRE: mean magnitude relative error, HGA: Hierarchical genetic algorithm, FRBFN: fuzzy radial basis function networks, HFRNN: Hesistant Fuzzy-rough Nearest-Neighbor, LRLT: Linear regression logarithmic transformation, JITL: Just in time learning, ELM: Extreme learning machine.

Modelling Outcome	Studies	Best model performance	Best baseline performance
Binary Classification:			
ICU LoS > 7 days	Wang et al.[28]	GRU-D: AUROC=0.71	RF: AUROC=0.764
HLoS > 7 days	Zhang et al.[18]	Fusion-CNN: AUROC=0.784	LR: AUROC=0.773
	Zebin, Rezvy, and Chaussalet[57]	Autoencoder+DNN: Accuracy=77.7%	RF: Accuracy=66.2%
	Alsinglawi et al.[50]	RF: AUROC=0.93	No baseline
ICU LoS > 3 days	Wang et al.[28]	GRU-D: AUROC=0.733	RF: AUROC=0.736
HLoS > 3 days	Alahmar, Mohammed, and Benlamri[17]	Stacked Ensemble: AUROC=0.81	No baseline
ICU LoS > 11.3 days	Gong et al.[83]	LR + NLP(with CUIs): AUROC = 0.77	LR + NLP(with shared item IDs): AUROC = 0.74
ICU LoS ≥ 8 days	Huang et al.[70]	CBFL-5 communities: AUROC=0.6512	FL: AUROC=0.636
HLoS > 5 days	Gentimis et al.[65]	median accuracy=79	No baseline%
HLoS > 41 days	Jalali et al.[54]	DNN: AUROC=0.94	No baseline
HLoS > 10 days	Ma et al.[55]	One-class JITL+ELM: AUROC=0.8510	One-class ELM: AUROC=0.7588
Composite outcome			
ICU LoS > 7 days & ihm	Weissman et al.[29]	GB(with text data): AUROC=0.89	GB(no text data): AUROC=0.83
ICU LoS > 21 days & ihm	Weissman et al.[29]	GB(with text data): AUROC=0.84	GB(no text data): AUROC=0.89

Continues on next page

Table 6 – *Best model performances clustered by outcome.*

Modelling Outcome	Studies	Best model performance	Best baseline performance
ICU LoS < 3 days, *ihm*, *cc*& *surg rec*	Che et al.[27]	GRU-D: AUROC=0.837	GRU-Simple: AUROC=0.825
Multi-Classification			
3 Classes:			
HLoS ≤ 10, 10 < HLoS ≤ 30, HLoS > 30 days	Harerimana, Kim, and Jang[56]	HAN + SMT: Micro AUROC = 0.8212	SAnD-Multi[75]: Micro AUROC=0.8190
ICU LoS ≤ 3, 3 < ICU LoS < 5, ICU LoS ≥ 5 days	Batista and Sanchez-Arias[48]	RF, SVM, Range RF: Accuracy=0.739	No baseline
ICU LoS ≤ 3, 3 < ICU LoS < 10, ICU LoS ≥ 10 days	Batista and Sanchez-Arias[48]	SVM, C5.0 GBT, C5.0 DT: Accuracy=0.927	No baseline
ICU LoS ≤ 1.8, 1.8 < ICU LoS < 5, ICU LoS ≥ 5 days	Batista and Sanchez-Arias[48]	Ranger RF: Accuracy=0.603	No baseline
1 ≤LoS ≤ 2, 3 ≤ LoS ≤ 5, LoS ≥ 6 days	Ramkumar et al.[68]	NB: AUROC = 0.8818	No baseline
1 ≤ LoS ≤ 3, 4 ≤ LoS ≤ 5, LoS > 5 days	Navarro et al.[67]	NB: AUROC = 0.7822	No baseline
1 ≤ HLoS ≤ 4, 5 ≤ HLoS ≤ 8, 9 ≤ HLoS ≤ 14 days	Alturki et al.[61]	SVM: Accuracy=87.48%	No baseline
4 Classes:			
1 ≤HLoS ≤ 3, 4 ≤ HLoS ≤ 6, HLoS ≥ 10 days	Karnuta et al.[53]	NB with adaptive boosting: AUROC=0.88	No baseline
9 Classes:			
ICU LoS of 1-7 days, 8 ≤ ICU LoS ≤ 14, ICU LoS > 14 days	Xu et al.[82]	RAIM-3: Accuracy=86.82%	CNN-RNN: Accuracy=85.34%
10 Classes: ICU LoS <1 day, ICU LoS of 1-7 days, 8 ≤ ICU LoS ≤ 14, ICU LoS > 14 days			
	Gao et al.[77]	Dr. Agent: kappa=0.3988	Dipole[103]: kappa=0.3965
	Harutyunyan et al.[76]	C+DS: kappa=0.451	LR: kappa=0.402
	Khadanga et al.[73]	MultiModal - 1DCNN(with text): kappa=0.453	LSTM(no text): kappa=0.438
Composite outcome			
ICU LoS in 10 classes, *ph*, *ihm* & *dc*	Song et al.[75]	SAnD-Multi: Kappa=0.429	LSTM: kappa=0.427
Regression			
HLoS			

Continues on next page

Table 6 – *Best model performances clustered by outcome.*

Modelling Outcome	Studies	Best model performance	Best baseline performance
	Purushotham et al.[58]	MMDL(FFN+GRU): MSE=53410.267 mins	Super Learner-II: MSE=54593.33 mins
	Zhang et al.[52]	GBM: refer to Figure 7	No baseline
	Balkan and Subbian[59]	RF: MAE=4.253 days	APACHE IVa: MAE=5.389 DAYS
	Sotoodeh and Ho[60]	HMM: RMSE=228.12	Lasso Regression: RMSE=234.19
	Mansouri, Noei, and Abadeh[47]	RF: $R^2 = 77.8\%$	No baseline
	Elrod et al.[63]	RF: MAE=3.73 days	No baseline
	Elbattah and Molloy[64]	RF: Relative absolute error $\approx$ 0.26	No baseline
ICU LoS			
	Sheikhalishahi, Balaraman, and Osmani[69]	BiLSTM: MAE=0.479 days	ANN: MAE=1.267 days
	Balkan and Subbian[59]	RF: MAE=2.010 days	APACHE IVa: MAE=2.491 DAYS
	Alsinglawi et al.[49]	Stacking regressor: MAE=1.92days	No baseline
	Rocheteau et al.[71]	LSTM-GAT, LSTM-MPNN: MAE=1.86 days	LSTM: MAE=1.95 days
Composite outcome			
ICU LoS & ihm	Rocheteau, Liò, and Hyland[14]	eICU data: TPC: MAE=1.55 days ; MIMIC-IV data: TPC: MAE=2.28 days	eICU data: Transformer: MAE=2.36 days ; MIMIC-IV data: TPC: MAE=3.61 days
Hospital high LoS outtliers (HHSOs)	Belderrar and Hazzab[66]	HGA+FRBFN: NICU: MMRE=0%, MICU: MMRE=2.13%, CCU: MMRE==6.31%, CSRU: MMRE=3.26%, SICU: MMRE=4.98%	No baseline
Clustering & Regression			
ICU LoS	Rouzbahman, Jovicic, and Chignell[74]	K-means+Linear R: MAE=7300 mins	Linear R: MAE=7474 mins
Regression & Classification			
HLOS	Wang et al.[62]	HFRNN: refer to paper on Figure 2	LRLT: refer to paper on Figure 2
Survival Analysis			

Continues on next page

Table 6 – *Best model performances clustered by outcome.*

Modelling Outcome	Studies	Best model performance	Best baseline performance
HLoS	Nemati, Ansary, and Nemati[51]	Stagewise GB: C-index=71.47	No baseline

3.6 Model Interpretation

As mentioned earlier, some models especially those belonging to the deep learning family, though very powerful for clinical data modelling outperforming most other models have a major drawback stemming from their inherent black-box nature. However, the importance of model interpretability can not be overlooked because it enables to understand which patient input parameters contributes the most to LoS predictions as well as the time value of the information (that is, whether the earliest or most recently available information contributes more to the prediction outcome). Some studies (not included in this review) showed that variables extracted on the fifth day of patient stay had more predictive value for predicting prolonged stays than those extracted on earlier days [109]. Table 7 summarizes the different tools used for model interpretability.

Table 7: Model Interpretability tools

Tool	**Description**	**Studies**
Regression	Linear regression models with logarithmic transformation (LRLT), and Cox regression model are used as feature selectors to select significant data and predictive attributes. Interpretation of logistic regression weights	[29, 62, 63, 83]
Results Visualisation	visualization of predictions over time for channel-wise LSTM with deep supervision to show areas of high probability prediction by the model and how they coincide with ground truth plots. Visualization via heatmaps of transition probabilities from hidden markov model	[60, 76]
Ablation Study	re-run models using different combinations of architectures like loss function, mask vectors, etc. or excluding some patient features to investigate their influence towards model prediction	[14, 18, 73]
Integrated Gradients	calculation of feature attributions towards prediction as we move from some baseline values to the input as explained in [110]	[14]
Attention mechanisms	mechanism by which to identify if a particular input or particular time step is important for the prediction by learning vectors of attention weights	[56, 71, 82]
Permutation	feature importance per feature by computing increase in prediction error when values of investigated feature are permuted and others kept constant	[64]
Feature Importance	obtained from XGBoost, random forest, SHAPLEY algorithms and GBoost	[47, 49, 57, 61]
Survival Analysis based model		[51]
Clustering		[70]
Analysis of decay weights	Analysis of decay curve weights to understand most imporant features and most predictive time points	[27]

4 Discussion and analysis of studies with respect to hospital management

As stated in the introduction, the goal of an intelligent LoS prediction system is to assist the hospital management team, the planning team more specifically to guarantee optimal patient flow, patient satisfaction, resource allocations and hospital performance and not to effectively discharge patients which is the role of the unit head. Thus, such a system should be specifically adapted to the hospital configuration and enable the management team to have a clear view on discharge times.

4.1 Main Contributions of Selected papers

4.1.1 Algorithms:

The foremost contribution of the various authors was the myriad of algorithms used and the adaptations performed towards LoS predictions for the specific patient populations and data sizes. With a sample size of 1592 patients suffering from heart failure, Alsinglawi et al. [49] showed that traditional boosted tree algorithms performed better than deep neural network architectures. This highlights the fact that different instances require the use of different algorithms and the most sophisticated are not always the ones to go with.

Several means were shown for improving prediction accuracy, from building meta-learners [17], or combining different data modalities via multi-modal architectures as explained in section 3.3.3, or the use of different model components like skip connections [14, 77] to allow each model layer to see the original data and reduce the effects of imputation, or the use of privacy-preserving data mechanisms through federated learning [70].

Another important contribution was the use of transformer networks and attention mechanisms to capture important time points of input features that contribute the most to prediction while providing interpretability [14, 56, 75, 82].

The use of sequence-to-sequence models that enable the continuous prediction and update of remaining LoS either as continuous or categorical outcomes during the patient stay also stands as an important contribution [14, 57, 73, 75, 76, 82].

The use of multi-centric data by a number of authors increases the generalization power of the developed models which could be indicative that different patient populations can be considered.

4.1.2 Clinical Outcome:

Though the prominent use of prediction models for LoS modelling is for capacity planning at medical unit or hospital level, the benefits to patient care come as a secondary advantage. That is, the stratification and prediction of patient stays as either short or extended, short, medium or long can enable the early identification of patients at risk of extended stays. This was carried out for ICU patients [18, 27–29, 48, 55–57, 65, 70, 73–76, 82, 83], diabetes patients [17, 61] and others [53, 67, 68]. For critically ill patients like ICU patients, an alert raised by the prediction model can be life saving and the counselling of patients at risk of extended stays can prepare them and their families psychologically during the hospitalization.

Another clinical advantage of some models was the continuous update of remaining ICU time predictions that enabled to take into account the continuous stream of incoming recording and adjust previously calculated predictions. The robustness of this model architecture was demonstrated in Rocheteau, Liò, and Hyland[14] when analyzing the improvement in accuracy and reliability of the predictions between admission and discharge for a single patient.

4.1.3 Feature Engineering:

Important contributions here include the indication of alternative techniques for categorical data encoding, which is very important for medical data where features like medications, procedures and diagnoses codes can have thousands of unique values [69, 71]. The use of entity embeddings instead of one-hot-encodings drastically reduced the embedding size of diagnoses codes while improving model performance in Sheikhalishahi, Balaraman, and Osmani[69].
Inclusion of experts in the loop using domain knowledge for creating new features also proved to be beneficial in Alturki et al.[61], where new features were created such as, service utilization computed as the sum of number of outpatients, emergencies and inpatients (in order to quantify the patientś usage of hospital services), number of medications and number of medication changes(to measure the number of times medication dosage changed).

4.2 Limitations of Selected papers

4.2.1 Hospital Management:

Problem Formulation Adequacy:
The majority of the studies approached the LoS prediction problem as a classification and not a regression problem, that is, by converting the original continuous outcome into buckets using different thresholds. These thresholds were not clinically motivated but were either chosen following previous studies, as the 85% percentile of the stays, median or mean number of days, etc. which might not be translated directly into a relevant parameter in a real hospital setting.
The major drawback in both the binary and multi-class (especially three classes) prediction formulation, was that they do not provide adequate information for effective patient scheduling. Predicting whether a patient will have a short, medium or long stay reduces the model utility for capacity planning.
Harutyunyan et al.[76] who initially considered the regression problem turned back to classification due to poor performance of regression models. This prompted the authors to perform a binary classification and an ordinal multi-classification study. In the binary classification study, the outcome was split in two classes for predicting prolonged LoS (for stays longer than seven days) or not, where as in the multi-classification study, the remaining hourly LoS was divided in ten buckets (one for extremely short visits of less than one day, seven day-long buckets for each day of the first week, and two outliers (per their definition of outlying observations) buckets - one for stays of over one week but less than two and one for stays of over two weeks. The latter was done to mimic hospitals custom practice according to Harutyunyan et al.[76]. As it is seen, this multi-classification setting appeared as a trick to reproduce the regression problem at least for the first eight classes where LoS is a discrete number. This appears more suitable for hospital management as it gives more precise information on patient discharge dates than binary or three-classes classification. The authors suggested that a reason for poor performance of regression models were the difficulties encountered by the prediction model in understanding the clear patterns for distinguishing between very close stays (e.g., a 3 day stay and a 4 day stay) or very distant stays (e.g., a 3 day stay and a 20 day stay).

Issues of Class Imbalance:
The simplification of the LoS problem from the original regression setting to a classification not only reduces model utility for hospital management but introduces issues like class imbalance, a situation in which a class label in the target variable is underrepresented in the data. In situations of class imbalance, the model predicts more accurately the instances truly belonging to the majority class than those belonging to the minority class that are more misclassified. Situating this regarding the LoS prediction problem, it means that more extreme stay patients (the minority but important class) will be predicted as short stay patients. This will then have negative consequences both for efficient hospital management as well as for the well-being

of patients since they might not be psychologically prepared for their truly long stay. In instances of class imbalance, the model might still report a high performance especially if metrics like accuracy or AUROC are used for model evaluation though the minority class will he highly misclassified as explained in Zolbanin et al.[15].

4.2.2 Limited Clinical Outcome and Data Diversity:

Most studies focused on the prediction of ICU LoS which is an important management tool for a critical care department but without extending this to the total hospital LoS or providing directions for extension. Because ICU patients are not directly discharged home but to acute care wards, optimal patient flow can only be ensured if the patient LoS is monitored in and out of ICU units till discharge.
The availability and use of ICU data both for ICU and hospital LoS predictions led to the construction of prediction models specifically adapted to the high density of ICU data without mention of the fit of less dense data from acute care wards where patients are not monitored continuously.

4.2.3 Patient Population and Model Input Features:

Since most freely and publicly available EHRs databases contain ICU patients, the algorithms developed were tested mostly on these populations without knowledge on how they will perform on acute care ward patients that have different data structures especially in terms of density. Another limitation in choosing a particular patient population is that, querying these specific populations required the use of discharge information, more specifically discharge diagnoses from the database limiting the deployment of such models in real-life settings since discharged diagnoses codes are not known at admission.
Given the complexity of the target, a limitation in some studies was the limited number of input features used for building the predictive model. Only eight input features were used in Navarro et al. [67] and Ramkumar et al.[68] and seven features in Sotoodeh and Ho [60]. Regarding input features, another limitation was the apriori choice of input features as mentioned in section 3.2 not guided by a team of medical experts but following similar works.

4.2.4 Model Evaluation and Explainability:

The mostly used metric in the regression setting was the coefficient of determination (R^2), a metric that does not objectively access model performance since its value increases with increase in the number of features in the model, thereby favoring most complex models.
As reported by some authors, the black-box nature of algorithms used and lack of explainability could greatly hinder their acceptance by practitioners [77].

5 Future research perspectives

The suggestions for further research into LoS predictions are summarized under the following points;

1. Problem Formulation: the predictive models built for LoS prediction should be such that they are useful to the hospital management. Predicting the number of days to discharge including the time of the day at which discharge is to occur might be an important outcome to consider or target to provide enough information for new patients admission and maintaining optimal patient flow; in other words, predicting exact discharge time with LoS in fractional days, that is, predicting 3.5 days to mean 3 days and discharge the fourth day at noon. This is important because it can enable to know if a new patient can be admitted on the same bed (including cleaning time) on the same day after another patient was discharged. In times of crisis like covid-19, such information would be valuable.

2. Data Population: The public release of acute ward data by hospitals should be encouraged to give the opportunity to researchers to develop models adaptable to highly sampled information from ICU from continuous patient monitory but at the same time less sampled data from acute care wards. This will enable the deployment of the developed algorithms in hospitals.

3. Model generalizability and Data Privacy: when data access makes it possible, datasets from different sources (medical institutions) should be considered to increase the ability of the developed models to be applicable to a broader range of patient populations. In such situations, the sensitivity of patient information from different sources should be emphasized to prevent data leakage. Techniques like federated learning as used in Huang et al.[70], differential privacy or a combination of both can be investigated to ensure that using trained models, original patient features cannot be recovered.

4. Informative Feature Engineering: the apriori selection of patient parameters to be used as model inputs can be very subjective. Moreover the features over which patients are monitored strongly depends on the individual patient conditions. Therefore, first, more informative features can be selected for modelling either by involving medical practitioners or using tools like word embeddings, that is, real-valued vectors representing words in the context in which they appear [111]. Secondly, more flexible models like graph neural networks that can handle varying input sizes per patient can be investigated such that most informative features at patient level are used.

5. Dealing with irregularly sampled time-series data: As explained in Section 3.3.3, the discrepancy in the interval of measurement between patient parameters and the combination of patient data of different structure creates irregular time intervals between lines of data per patient, a situation that worsens when all patients are combined to create a single tabular dataframe. So far, the mostly investigated technique for ICU patients was hourly re-sampling which offers a limitation in that the granularity in data information is lost. This might have a more pronounced effect for acute care ward patients as data are more scanty. The use of model structures capable of capturing this irregularity can be investigated and features included in the model to capture this information.

6. Dealing with real-time predictions when new data becomes available: most consulted authors updated trained models hourly thereby increasing the sequence of past information as the patient stays within hospital premises. Suggested alternatives could be either to use a moving window such that only recent information is used for model update or the model is only re-trained when triggered by a particular patient condition.

7. Model Interpretability: though the best performing algorithms are neural-network based, that is, black-box models, the question of interpretability can not be overlooked. More efforts can be done to provide the parameters driving the predictions both at the model level and at the individual patient level. Feature attribution algorithms like integrated gradients or SHAPLEY additive algorithms [112] can be considered. The latter provides among other information, overall feature importance, overall feature importance combined with feature effects such that it is possible to visualize the impact of feature values per feature towards model prediction and interaction effects between multiple features.

8. Model Evaluation and Clinical Validation: As discussed before, one of the limitations of the consulted studies is the use of metrics that do not adequately express actual performance. In the regression setting, more intuitive metrics like the mean absolute error (MAE) or mean absolute percentage error (MAPE) can be used. For classification problems and in the presence of class imbalance, metrics like the area under the precision-recall (AUPRC) curve should be preferred [113] over the AUROC since the former uses the precision measure (number of correctly classified samples from the minority class)

instead of the false positive rate which is less informative in this situation because there are more correctly classified samples from the majority class (more true negatives).

To further increase the transparency in the models it is important that medical practitioners are involved in all steps of the research so they can understand and validate each stage from feature engineering through prediction and model interpretability. The use of prospective data on which the models are deployed will further validate the models. Ultimately, the models should be evaluated by studying the effect of their implementation on actual patient outcomes and throughput.

6 Conclusion

LoS, the duration of stay of patients in a single care episode is an important measure of hospital efficiency. Early prediction and continuous monitoring of patients' LoS during hospitalization is necessary both for improved hospital managements as well as patient satisfaction. By so doing, medical institutions will be able to efficiently allocate resources at any time and be better prepared during critical periods like pandemics.

In the current work, we provided a systematic and comprehensive review of recent studies on the prediction of LoS using machine learning and deep learning techniques. Thirty-eight research papers published between January 2016 and April 2021 retrieved from six unique repositories were thoroughly reviewed and summarized under five aspects. Concerning LoS predictions, most authors considered the classification approach in order to reduce modelling difficulties though incurring utility loss by hospital management. This classification scheme also brought up a number of challenges among which class imbalance that heavily mislead model accuracy. Other authors adapted the multi-classification schemes so that these can provide more granular information to the management team of the hospital. The mostly used publicly available dataset were MIMIC-III [72] and eICU-CRD [79] consisting of intensive care patients. For data modelling, linear and logistic regression models were mostly used as baselines. Random forest had the best performance on classification tasks but required hand-crafted features extracted from time-series data. Sequence-to-sequence models that inherently take into account the time dimension in sequential data were the mostly used type of neural networks. Also, the latter showed high abilities for overcoming existing challenges in EHR modelling. More recent sequential architectures like transformer networks and temporal convolution networks also demonstrated high prediction capabilities while providing interpretability in the case of the former. Some authors list out a number of factors that greatly influenced either positively or negatively the LoS and could be used to build a risk factor assessment tool for prolonged LoS patients. This is useful especially if it can considerably reduce input space for real-time LoS predictions while maintaining high prediction accuracy. The lack of interpretability and validation of the models by clinical experts could be a hindrance to their acceptance by medical practitioners for implementation in a real hospital setting. Authors from recent years mostly employed Deep neural network based models taking into account the time-dimension in the data compared to those from earlier years. The latter mostly used early and static patient data to which traditional machine learning methods were applied. Finally, some research perspectives are proposed to broaden the research area.

Funding

This work was supported by KU Leuven: Research Fund (projects C16/15/059, C3/19/053, C24/18/022, C3/20/117, C3I-21-00316), Industrial Research Fund (Fellowships 13-0260, IOFm/16/004, IOFm/20/002) and several Leuven Research and Development bilateral industrial projects; Flemish Government Agencies: FWO: EOS Project no G0F6718N (SeLMA), SBO project S005319N, Infrastructure project I013218N, TBM Project T001919N; PhD Grants (SB/1SA1319N, SB/1S93918, SB/1S1319N), EWI: the Flanders AI Re-

search Program VLAIO: CSBO (HBC.2021.0076) Baekeland PhD (HBC.20192204) and Innovation mandate (HBC.2019.2209) European Commission: European Research Council under the European Unionâs Horizon 2020 research and innovation programme (ERC Adv. Grant grant agreement No 885682); Other funding: Foundation âKom op tegen Kankerâ, CM (Christelijke Mutualiteit)

Conflict of interest

The authors declare no conflict of interest.

Appendix

A Datasets

List of currently available pipelines that facilitate data extraction in Table 8.

B Model Peformance as reported by various authors

Model performance as evaluated using performance metrics on developed and implemented baseline models are given in Table 9.

Table 8: Available pipelines for data access

Name of Dataset	Pipelines	Studies
MIMIC-II (Multiparameter Intelligent Monitoring in Intensive Care)	Original pipeline provided by data host at https://archive.physionet.org/mimic2/mimic2_access.shtml	[74]
MIMIC-III (Medical Information Mart for Intensive Care III)	Original pipeline provided by data host at: https://physionet.org/content/mimiciii/1.4/	[18, 29, 47, 50, 65, 66, 83]
	MIMI-Extract[28]: open source pipeline data extraction from MIMIC-III v1.4. Outlier detection and correction, hourly sampling, measurement unit standardization are performed on extracted features. Aggregation of semantically similar features is done.	[28]
	Pipeline from [58]: available code to extract	[58]
	Pipeline from [76]: available code to extract data on 17 features	[73, 75, 76, 82]
MIMIC-IV	Pipeline from [14]: available code to extract data for 69,609 ICU stays from 50,042 patients. 172 time series from the chartevents and labevents and 12 static features.	[14]
MIMIC-III Waveform database		[82]
eICU (Philips eICU Collaborative Research Database)	Pipeline from [69]:	[69]
	Pipeline from [14]: available to extract data of 118,535 adult patients and 146,671 ICU stays. 87 time series features, 17 static features and diagnoses.	[14]
Physionet 2012 Challenge	Data available via url provided in Table 2	[60]
Heritage Health prize data	Data available via url provided in Table 2	[52]
UCI diabetes dataset	available for download on host page	[17, 61]
Irish Hip Fracture data	Data available via url provided in Table 2	[64]
NewYork State Department ofHealthâs Statewide Planning and Research Cooperative System (SPARCS)	Data available via url provided in Table 2	[53, 67, 68]
PHN SVR trial: Pediatric Heart Network (PHN) that conducted the Single Ventricle Reconstruction (SVR) Trial	Original pipeline provided by host at: https://www.pediatricheartnetwork.org/ForResearchers/PHNPublicUseDatasets/SingleVentricleReconstruction-SVR	[54]
Web-based international burn registry of German- speaking countries	available for download on host page	[63]
PMSI MCO: Program for medicalization of information systems in medicine, surgery, obstetrics and odontology	Data available on request from provided url in Table 2	[62]
Open-access dataset	Data available via url provided in Table 2	[51]

Table 9: Prediction outcome and model performance. GCS: Glasgow Coma Score, C.I: Confidence Interval, RAIM: Recurrent Attentive and Intensive Model, FL: federated learning, CBFL: Community-Based FL, CL: Centralized Learning, SAnD: Simply Attend and Diagnose, MMDL: Multi-Modal Deep Learning, SoI: Severity of Illness, RoM: Risk of Mortality, HAN: Hierarchical Attention Network, SGD: Stochastic Gradient Descent, SVR: Support Vector Regression, ELM: Extreme Learning Machine, HMM: Hidden Markov Model, NIDiasABP: Non-Invasive Diastolic Arterial Blood Pressure, NISysABP: Non-Invasive Systolic Arterial Blood Pressure, MAP: Mean Arterial Blood Pressure, MPNN: Message Passing Neural Networks, GAT: Graph Attention Networks, UMLS: Unified Medical Language System, ED: Emergency Department, HFRNN: hesitant-fuzzy nearest neighbor algorithm, LRLT: Linear regression logarithmic transformation

Study	Outcome (only LoS)	Model(s)	Independent variables	Performance
[57]	HLoS> 7days	Autoencoder & MLP with 3 layers	demographics, diagnosis codes, first 24hours values of temperature, blood pressure, heart rate	Accuracy=77.7%
[49]	ICU LoS regression	RF, GBR, MLP with 3 layers, Stacking Regressor (SR)	Age, weight, gender, temperature, heart rate, respiratory rate, pulse oximetry	RF: R^2 = 0.8, MAE=1.98, time=0.95seconds ; GBR: R^2 = 0.81, MAE=2, time=0.85seconds; SR: R^2 = 0.81, MAE=1.92, time=11.84sec; DNN: R^2 = 0.77, MAE=2.3, time=4.115seconds
[76]	10-class classification of remaining ICU LoS at each hour; ICU LoS <1 day, 1-7: 7 classes for each day of the week, over one week but less than two, stays over two weeks(after first four hours of ICU stay)	LR, standard LSTM (S), channel-wise LSTM (C), Multi-task standard LSTM (MS), Multi-task channel wise LSTM (MC), Deep supervision (DS)	capillary refill rate, diastolic blood pressure, fraction inspired oxygen, GCS eye opening, GCS motor response, GCS total, GCS verbal response, glucose, heart rate, height, mean blood pressure, oxygen saturation, respiratory rate, systolic blood pressure, temperature, weight, pH	metric value (95% CI 1000 boostraps) LR: kappa=0.42(0.401, 0.404); S: kappa=0.438(0.436, 0.440); S+DS: kappa=0.431(0.430T, 0.433); C: kappa=0.442(0.440, 0.444); C+DS: kappa=0.451(0.499, 0.453); MS: kappa=0.450(0.449, 0.452); MC: kappa=0.450(0.448, 0.451)
[82]	9-class ICU LoS, ICU LoS 1-7 days, >8 days and up to two weeks of stay, over two weeks of stay	CNN, CNN-RNN, CNN-AttRNN, RAIM-0, RAIM-1, RAIM-2, RAIM-3	discrete clinical records, high-density waveforms, numerical vital signs	CNN(ECG): kappa=0.77, Accuracy=82.16%; CNN-RNN: kappa=0.80, Accuracy=85.34%; CNN-AttRNN: kappa=0.82, Accuracy=84.89%; RAIM-0: kappa=0.81, Accuracy=85.84%; RAIM-1: kappa=0.82, Accuracy=86.74%; RAIM-2: kappa=0.82, Accuracy=85.21%; RAIM-3: kappa=0.83, Accuracy=86.82%
[14]	Remaining hourly ICU LoS regression	Temporal Pointwise Convolution network (TPC) with 9 layers	diagnoses, static features, time series	10 runs of models ±95% CI: eICU: LSTM: MAD = 2.39±0.00, MAPE=118.2 ± 1.1, MSE=26.9 ± 0.1, MSLE=1.47 ± 0.01, R^2 = 0.09 ± 0.00, kappa=0.28 ± 0.00 ; Channel-wise LSTM: MAD = 2.37±0.00, MAPE=114.5 ± 0.4, MSE=26.6 ± 0.1, MSLE=1.43 ± 0.01,R^2 = 0.10 ± 0.00, kappa=0.30 ± 0.00 ; Transformer: MAD = 2.36±0.00, MAPE=114.1 ± 0.6, MSE=26.7 ± 0.1, MSLE=1.43 ± 0.01,R^2 = 0.09 ± 0.00, kappa=0.30 ± 0.00 ; TPC: MAD = 1.78±0.02, MAPE=63.5 ± 4.3, MSE=21.7 ± 0.5, MSLE=0.70 ± 0.03,R^2 = 0.27 ± 0.02, kappa=0.58 ± 0.01 ;

Continues on next page

Table 9 – *Continued from previous page*

Study	Outcome (only LoS)	Model(s)	Independent variables	Performance
[52]	HLoS regression	RF, GBR, Linear R	139 features extracted from claims data, drugs and labs	refer to paper on Figure 7.
[65]	HLoS>5 days	MLP with 2 layers	admissions, cost events, services, ICU admission and discharge units, demographics, procedures, diagnosis	100 runs (C.I): Accuracy = 79% (75.3%, 82.3%)
[17]	HLoS>3days	Stacked ensemble	demographics, diagnoses, admission entry, procedures, medications, laboratory, number of emergency visits, discharge disposition	9-fold CV: AUROC=0.81
[61]	3-class HLoS, short:[1-4]days, medium:[5-8]days, long:[9-14]days	LR, RF, SVM (RBF kernel), XGBoost, KNN(1-neighbor)	demographics, diagnoses, admission entry, procedures, medications, laboratory, number of emergency visits,etc.	RF: Accuracy=79.84%, precision=81.23%, recall=79.84%, F1-score=79.49%; LR: Accuracy=54.76%, precision=53.71%, recall=54.76%, F1-score=54.03%; XGBoost: Accuracy=81.02%, precision=81.04%, recall=80.95%, F1-score=80.81%; SVM: Accuracy=87.48%, precision=87.65%, recall=87.48%, F1-score=87.39%; KNN: Accuracy=85.46%, precision=86.79%, recall=85.46%, F1-score=84.82%
[73]	Similar to [76]	1-D CNN, LSTM	time-series patient data from [76] and clinical notes	5 runs: MultiModal - 1DCNN: kappa=0.453
[70]	ICU LoS $\geq$ 8 days	CBFL, FL, CL	1399 binary drug features administered during first 48hours of ICU stay	CL: AUROC=61.89%, AUC-PR=6.2%; FL: AUROC=62.12%, AUC-PR=7.2%; CBFL(with 5 communities): AUROC=64%, AUC-PR=8.22%
[75]	Similar to [76]	LR, LSTM, SAnD, LSTM-Multi, SAnD-Multi		LR: kappa=0.402, MSE=63385, MAPE=573.5, LSTM: kappa=0.427, MSE=42165, MAPE=235.9, SAnD: kappa=0.429, MSE=40373, MAPE=167.3, LSTM-Multi: kappa=0.426, MSE=42131, MAPE=188.5, SAnD-Multi: kappa=0.429, MSE=39918, MAPE=157.8
[58]	HLoS in minutes	Super learners (ensemble of multiple models), FFN, GRU, Multi-modal DL(MMDL)	Feature set A, Feature set B and Feature set C	**Super Learner II on Feature set C**: On first 24-h data: MSE= 54608.1099 $\pm$ 2923.9972, On First 48-h data: MSE=54400.5845 $\pm$ 1582.4523; **FFN on Feature set C**: On first 24-h data; MSE= 53410.0918 $\pm$ 3207.9849, On first 48-h data: MSE= 52642.6508 $\pm$ 4373.4239y; **GRU on Feature set C**: On first 24-h data: MSE=48702.7641 $\pm$ 3768.5154, On first 48-h data: MSE= 49556.8024 $\pm$ 3794.0471; **MMDL On Feature set C**: On first 24-h data: 36338.2015 $\pm$ 2672.3832, On first 48-h data: MSE = 36924.2312 $\pm$ 3566.4318
[67]	$1 \leq$ LoS ≤ 3, $4 \leq$ LoS ≤ 5, LoS > 5	Naive Bayes(NB)	age group, Charlson comorbidity score, ethnicity, gender, patient disposition, admission type, RoM, SoI	5-fold CV: AUROC = 78.22%, Accuracy = 87.39%
[47]	HLoS regression	RF, XGBoost, KNN, Linear R, GBR	features extracted from demographics and diagnoses codes	RF: R^2 = 77.80%; XGBoost: R^2 = 76.08%; KNN: R^2 = 73.06%; Linear R: R^2 = 64.66%; GBR: R^2 = 76.51%

Continues on next page

Table 9 – *Continued from previous page*

Study	Outcome (only LoS)	Model(s)	Independent variables	Performance
[27]	Composite outcome including ICU LoS $<$ 3 days	GRU-D	33 features extracted from time series vital signs and labs	Average AUROC (mean$\pm$ std) = 0.8370 ± 0.012
[69]	Remaining ICU LoS in days hourly	Linear R, ANN with 1 hidden layer, BiLSTM	Heart rate, mean arterial pressure, diastolic blood pressure, systolic blood pressure, oxygen, respiratory rate, temperature, glucose, pH, height, weight, age, admission diagnosis, ethnicity, gender, GCS Total, GCS Eyes, GCS Motor, GCS Verbal	5-fold CV: Linear R (with Numerical and categorical features): $R^2 = 0.024 \pm 0.001$, MAE(in days)=1.29 ± 0.008; ANN(with Numerical and categorical features): $R^2 = 0.048 \pm 0.003$, MAE(in days)=1.267 ± 0.014 ;BiLSTM(with numerical and categorical features): $R^2 = 0.643 \pm 0.042$, MAE(in days)=0.532 ± 0.033; BiLSTM(without categorical features): $R^2 = 0.610 \pm 0.042$, MAE(in days)=0.479 ± 0.016
[74]	ICU LoS regression in minutes	K-Means, Linear R	diagnoses codes, labs	MAE = 7474 on raw data, MAE = 7300 on clustered data
[53]	HLoS in days: 1 $\leq$LoS $\leq$ 3, 4 $\leq$ LoS $\leq$ 6, LoS $\geq$ 10	Naive Bayes(NB) with adaptive boosting	age group, sex, ethnicity, race, admission SoI and RoM, admission type	10-fold CV: AUROC = 88%, Accuracy = 76.5%
[68]	HLoS in days: 1 $\leq$LoS $\leq$ 2, 3 $\leq$ LoS $\leq$ 5, LoS $\geq$ 6	Naive Bayes(NB)	age group, Charlson commorbidity, ethnicity, gender, discharge disposition, RoM, SoI, admission type	AUROC = 0.8818, Accuracy = 83.08%
[56]	HLoS $\leq$ 10, 10 $<$ HLoS $\leq$ 30, HLoS $>$ 30	HAN + SMT, SMOOTEBOOST, H.attention	age, marital status, insurance type, smoking history, admission type, transfers, previous LoS, time between subsequent admissions, ICD 9-diagnoses and procedure codes, free text diagnoses	HAN+SMT: Macro AUROC = 0.7625; SMOOTEBOOST: Macro AUROC = 0.7103; H.attention: Macro AUROC = 0.6811
[48]	ICU LoS in days. Exp 1: LoS $\leq$ 3, 3 $<$ LoS $<$ 5, LoS $\geq$ 5; Exp 2: LoS $\leq$ 3, 3 $<$ LoS $<$ 10, LoS $\geq$ 10; Exp 3: LoS $\leq$ 1.8, 1.8 $<$ LoS $<$ 5, LoS $\geq$ 5	SVM, LR, RF, XGBoost, Linear R	demographics, ICD diagnoses groups	Refer to paper on Figure 2
[83]	ICU LoS $>$ 11.3 days	L2-regularized LR & NLP	clinical concepts based on UMLS concept unique identifiers (CUIs) extracted from labs and chart events	5-fold CV: CUIs: AUROC = 0.77(0.76, 0.78); All items IDs: AUROC = 0.70(0.69, 0.71); Shared Item IDS: AUROC = 0.74(0.73, 0.75)
[64]	HLoS regression	RF, boosted DT, DNN, Linear R	source hospital, discharge code, residence area, admission source, transfer, age, admission trauma type, pre-fracture mobility, fragility history, multi-rehabilitation assessment, admission type, sex, discharge status, ICD-10 diagnoses codes, admission via ED, fracture type, specialist falls assessment	10-fold CV: RF: RAE $\approx$ 0.26; boosted DT: RAE $\approx$ 0.34; DNN: RAE $\approx$ 0.55; Linear R: RAE $\approx$ 0.0.93

Continues on next page

Table 9 – *Continued from previous page*

Study	Outcome (only LoS)	Model(s)	Independent variables	Performance
[77]	Similar to [76]	Dr. Agent, LR, MLP, GRU, Dipole, T-LSTM	time-series from [76]	Dr. Agent: Kappa=0.3988(0.3955-0.4022), MSE=38446(38084-38816), MAPE = 188.8(186.9-190.5); LR: Kappa=0.3504(0.3477-0.3534), MSE=47928(47557-48335), MAPE=237.4(234.9-239.7); MLP: Kappa=0.3577(0.3545-0.3599), MSE=46804(46227-47364), MAPE=230.0(227.4-233.1); GRU: Kappa=0.3864(0.3831-0.3898), MSE=39310(38976-39651), MAPE=187.2(185.4-189.2); Dipole: Kappa=0.3965(0.3935-0.3996), MSE=38971(38627-39321), MAPE=193.2(191.5-194.6); T-LSTM: Kappa=0.3872(0.3843-0.3902), MSE=40399(40042-40756), MAPE=189.7(188.6-191.0)
[50]	ICU LoS > 7 days	LR, RF, SGD, KNN, DT, Gaussian NB, SVM	37 features extracted from gemographics, diagnoses and procedure codes, transfers, labs, microlabs	5 fold-CV: LR: AUROC= 89%; RF: AUROC=93%; SGD: AUROC=82%; KNN: AUROC=73%; DT: AUROC=90%; Gaussian NB: AUROC=80%; SVM: AUROC=83%
[59]	HLoS, ICU LoS	RF	diagnoses codes, demographics	ICU LoS: Explained Variance = 0.162, MAE=2.010, R^2 = 0.016; Hospital LoS: Explained Varaince = 0.156, MAE=4.253, R^2 = 0.156
[54]	HLoS > 41 days	Ridge LR, DT, RF, GBT, DNN	age, sex, race, gestation age at birth, poverty score, preoperative variables, intraoperative variables	5-fold CV: Ridge LR: AUROC=0.54±0.07, Accuracy=0.63 ± 0.06; DT: AUROC=0.59 ± 0.05, Accuracy=0.53 ± 0.04; RF: AUROC=0.67 ± 0.03, Accuracy=0.61 ± 0.03; GBT: AUROC=0.88 ± 0.03, Accuracy=0.82 ± 0.03; DNN: AUROC=0.94 ± 0.04, Accuracy=0.85 ± 0.03
[18]	HLoS >7 days	LR, RF, Fusion-CNN, Fusion-LSTM	Static demographic information (S), Temporal vital signs(T), unstructured notes (U)	5 runs with different data splits (95%C.I): LR(U+T+S): F1-score=0.703 (0.699, 0.707), AUROC=0.773 (0.770, 0.776), AUPRC=0.642 (0.637, 0.647); RF(U+T+S): F1-score=0.537 (0.533, 0.541), AUROC=0.718 (0.714, 0.722), AUPRC=0.597 (0.591, 0.603); Fusion-CNN(U+T+S): F1-score=0.725 (0.718, 0.732), AUROC=0.784 (0.781, 0.787), AUPRC=0.662 (0.658, 0.666); Fusion-LSTM(U+T+S): F1-score=0.716 (0.711, 0.721), AUROC=0.778 (0.776, 0.780), AUPRC=0.660 (0.657, 0.663)
[63]	HLoS in days	Rule of thumb (RT), Linear R	TBSA by degree of burn, age, gender, inhalation trauma, injury, cause of injury, ventilation and treatment in burn ICU	20-fold CV: RT: MAE = 4.04, MSE = 45.51; Linear R: MAE = 3.81, MSE = 35.55; RF: MAE = 3.73, MSE = 35.85

Continues on next page

Table 9 – *Continued from previous page*

Study	Outcome (only LoS)	Model(s)	Independent variables	Performance
[66]	Hospital high LoS outliers	Hierarchical genetic algorithm & fuzzy radial basis function networks	28 features extracted from demography, admission information, labs, microlabs, medications, output fluids and procedure codes	Neonatatl ICU: MMRE = 0, Pred(q) = 100%; Medical ICU: MMRE = 2.13, Pred(q) = 94.69%; Coronary care unit: MMRE = 6.31, Pred(q) = 90.23%; Cardiac Surgery Recovery unit: MMRE = 3.26, Pred(q) = 92.33%; Surgical ICU: MMRE = 4.98, Pred(q) = 91.30%
[60]	HLoS regression	Gaussian HMM, LASSO regression, Ridge regression, Poisson regression, Negative binomial regression, Linear SVR, RBF SVR, SAPS-II	GCS, temperature, heart rate, WBC count, serum glucose, urine output, NIDiasABP	Gaussian HMM: RMSE=225.05; LASSO regression: RMSE= 234.19; Ridge regression: RMSE= 236.91; Poisson regression: RMSE= 235.26; Negative Binomial: RMSE= 235.01; Linear SVR: RMSE= 237.99; RBF SVR: RMSE= 236.97; SAPS-II: RMSE=225.05
[29]	Exp 1: composite outcome of ICU LoS $\geq$ 7 days and death, Exp 2: composite outcome of ICU LoS $\geq$ 21	LR, GBM, RF, Elastic Net regression	age, urine output, modified Elixhauser score, most severe value within first 48 hours of creatinine, WBC count, highest value of total bilirubin, heart rate, temperature, lowest value of platelet count, arterial partial pressure of oxygem, GCS, systolic blood pressure, gender, admission type, presence within 48 hours of mechanical ventilation, ICU admission and cardiac arrest	5 times 10-fold CV with inclusion of clinical notes: Exp 1: LR: AUROC=0.86(0.85-0.88); GBM: AUROC=0.89(0.88-0.90); RF: AUROC=0.88(0.87-0.89); Elastic Net regression: AUROC=0.88(0.86-0.89); Exp 2: LR: AUROC=0.85(0.83-0.86); GBM: AUROC=0.89(0.88-0.90); RF: AUROC=0.87(0.85-0.88); Elastic Net regression: AUROC=0.87(0.86-0.88)
[55]	HLoS > 10 days	Just-In Time Learning (JITL) & One-class ELM, One-class SVM, Binary ELM	Temperature, heart rate, urine, pH, respiratory rate, GCS, fractional inspired O_2, GCS, partial pressure CO_2, Invasive MAP, Invasive SysABP, Invasive DiasABP, NIMAP, NIDiasABP, NISysABP, mechanical ventilation	JITL & One-class ELM: AUROC=85.10, Accuracy=82%; Binary ELM: AUROC=49.73, Accuracy=50.25%; Binary JITL-ELM: AUROC=20.14, Accuracy=25.0%; One-Class ELM: AUROC=75.88, Accuracy=53.75%; One-class JITL-ELM: AUROC=85.10, Accuracy=82%; One-class SVM: AUROC=46.47, Accuracy=52.0%
[51]	Discharge time prediction	IPC Ridge, CoxPH, Coxnet, StagewiseGB, ComponentwiseGB, Fast SVM, Fast Kernel SVM	age, sex, time until discharge	IPC Ridge: C-index=49.05; CoxPH: C-index=70.03; Coxnet: C-index=70.72; StagewiseGB: C-index=71.47; ComponentwiseGB: C-index=70.60; Fast SVM: C-index=70.65; Fast Kernel SVM: C-index=61.05
[28]	Exp 1: ICU LoS > 3 days, Exp 2: ICU LoS > 7 days	LR, RF, GRU-D	gender, age buckets, ethnicity, ICU type, admission type, time of the day, time-varying labs and vitals	Exp 1: LR: AUROC=71.6, AUPRC=65.1, Accuracy=68.6%; RF: AUROC=73.6, AUPRC=68.5, Accuracy=69.5%; GRU-D: AUROC=73.3, AUPRC=68.5, Accuracy=68.3%; Exp 2: LR: AUROC=72.4, AUPRC=18.5, Accuracy=91.9%; RF: AUROC=76.4, AUPRC=19.5, Accuracy=91.9%; GRU-D: AUROC=71.0, AUPRC=17.9, Accuracy=91.2%

Continues on next page

Table 9 – *Continued from previous page*

Study	Outcome (only LoS)	Model(s)	Independent variables	Performance
[71]	ICU LoS regression	LSTM*(no diagnoses), LSTM, LSTM-SAGE, LSTM-GAT, LSTM-MPNN	38 features extracted from demographics, admission information, diagnoses, time-varying labs and chart events	C.I from 15 independent training runs: LSTM*(no diagnoses): MAD=1.97 $\pm$ 0.01, MSLE=0.398 $\pm$ 0.004, $Kappa$ = 0.224$\pm$0.006; LSTM: MAD=1.95$\pm$0.01, MSLE=0.382$\pm$0.001, $Kappa$ = 0.245 $\pm$ 0.006; LSTM-SAGE: MAD=1.87 $\pm$ 0.00, MSLE=0.377 $\pm$ 0.002, $Kappa$ = 0.237 $\pm$ 0.006; LSTM-GAT: MAD=1.86 $\pm$ 0.00, MSLE=0.371 $\pm$ 0.001, $Kappa$ = 0.258 $\pm$ 0.004; LSTM-MPNN: MAD=1.86 $\pm$ 0.01, MSLE=0.369 $\pm$ 0.001, $Kappa = 0.261 \pm 0.005$
Wang et al.	HLoS classification & regression	HFRNN, LRLT	principial diagnosis, age, gender, emergency, medical unit	refer to paper on Figure 2.

References

[1] Olubanke Babalola. "Consumers and their demand for healthcare". In: *J Health Med Econ* 3.1 (2017), p. 6.

[2] Sundiatu Dixon-Fyle and Thomas Kowallik. "Engaging consumers to manage health care demand". In: *McKinsey Quarterly, Januar* (2010).

[3] Ruth McCabe et al. "Adapting hospital capacity to meet changing demands during the COVID-19 pandemic". In: *BMC medicine* 18.1 (2020), pp. 1–12.

[4] Eurostat. *Ageing Europe - statistics on population developments - Statistics Explained*. 2019. URL: https://ec.europa.eu/eurostat/statistics-explained/index.php/Ageing_Europe_-_statistics_on_population_developments (visited on 12/01/2020).

[5] Jean-Pierre Michel and Fiona Ecarnot. *The shortage of skilled workers in Europe: its impact on geriatric medicine*. 2020.

[6] Stephen Swensen, Steven Strongwater, and Namita Seth Mohta. "Leadership survey: immunization against burnout". In: *NEJM Catalyst* 4.2 (2018).

[7] OECD iLibrary. *Health at a Glance: Europe 2020 : State of Health in the EU Cycle*. 2020. URL: https://www.oecd-ilibrary.org/sites/c762d8be-en/index.html?itemId=/content/component/c762d8be-en (visited on 06/22/2021).

[8] Segen's Medical Dictionary. *length of stay*. 2011. URL: https://medical-dictionary.thefreedictionary.com/length+of+stay (visited on 03/03/2021).

[9] Sean Barnes et al. "Real-time prediction of inpatient length of stay for discharge prioritization". In: *Journal of the American Medical Informatics Association* 23.e1 (2016), e2–e10.

[10] Gregory Mak et al. "Physiciansâ ability to predict hospital length of stay for patients admitted to the hospital from the emergency department". In: *Emergency medicine international* 2012 (2012).

[11] Antonio Paulo Nassar Jr and Pedro Caruso. "ICU physicians are unable to accurately predict length of stay at admission: a prospective study". In: *International Journal for Quality in Health Care* 28.1 (2016), pp. 99–103.

[12] Fábio Gusmão Vicente et al. "Can the experienced ICU physician predict ICU length of stay and outcome better than less experienced colleagues?" In: *Intensive care medicine* 30.4 (2004), pp. 655–659.

[13] MIT Critical Data. *Secondary analysis of electronic health records*. Springer Nature, 2016.

[14] Emma Rocheteau, Pietro Liò, and Stephanie Hyland. "Temporal pointwise convolutional networks for length of stay prediction in the intensive care unit". In: *Proceedings of the Conference on Health, Inference, and Learning*. 2021, pp. 58–68.

[15] Hamed M Zolbanin et al. "Data analytics for the sustainable use of resources in hospitals: Predicting the length of stay for patients with chronic diseases". In: *Information & Management* (2020), p. 103282.

[16] Jingyi Wu et al. "Development of a scoring tool for predicting prolonged length of hospital stay in peritoneal dialysis patients through data mining". In: *Annals of Translational Medicine* 8.21 (2020).

[17] Ayman Alahmar, Emad Mohammed, and Rachid Benlamri. "Application of data mining techniques to predict the length of stay of hospitalized patients with diabetes". In: *2018 4th International Conference on Big Data Innovations and Applications (Innovate-Data)*. IEEE. 2018, pp. 38–43.

[18] Dongdong Zhang et al. "Combining structured and unstructured data for predictive models: a deep learning approach". In: *BMC medical informatics and decision making* 20.1 (2020), pp. 1–11.

[19] Jaret M Karnuta et al. “The value of artificial neural networks for predicting length of stay, discharge disposition, and inpatient costs after anatomic and reverse shoulder arthroplasty”. In: *Journal of Shoulder and Elbow Surgery* 29.11 (2020), pp. 2385–2394.

[20] Aditya V Karhade et al. “Development of predictive algorithms for length of stay greater than one day after one-or two-level anterior cervical discectomy and fusion”. In: *Seminars in Spine Surgery*. Elsevier. 2021, p. 100874.

[21] Aya Awad, Mohamed Bader-El-Den, and James McNicholas. “Patient length of stay and mortality prediction: a survey”. In: *Health services management research* 30.2 (2017), pp. 105–120.

[22] Stephen Bacchi et al. “Machine Learning in the Prediction of Medical Inpatient Length of Stay”. In: *Internal Medicine Journal* (2020).

[23] Alireza Atashi et al. “Models to predict length of stay in the Intensive Care Unit after coronary artery bypass grafting: a systematic review.” In: *The Journal of cardiovascular surgery* 59.3 (2018), pp. 471–482.

[24] Ahmed Almashrafi, Mustafa Elmontsri, and Paul Aylin. “Systematic review of factors influencing length of stay in ICU after adult cardiac surgery”. In: *BMC health services research* 16.1 (2016), pp. 1–12.

[25] Li Luo et al. “Predicting intracerebral hemorrhage patientsâ length-of-stay probability distribution based on demographic, clinical, admission diagnosis, and surgery information”. In: *Journal of healthcare engineering* 2019 (2019).

[26] Staffs Keele et al. *Guidelines for performing systematic literature reviews in software engineering*. Tech. rep. Citeseer, 2007.

[27] Zhengping Che et al. “Recurrent neural networks for multivariate time series with missing values”. In: *Scientific reports* 8.1 (2018), pp. 1–12.

[28] Shirly Wang et al. “Mimic-extract: A data extraction, preprocessing, and representation pipeline for mimic-iii”. In: *Proceedings of the ACM Conference on Health, Inference, and Learning*. 2020, pp. 222–235.

[29] Gary E Weissman et al. “Inclusion of unstructured clinical text improves early prediction of death or prolonged ICU stay”. In: *Critical care medicine* 46.7 (2018), p. 1125.

[30] Matthew S Durstenfeld et al. “Physician predictions of length of stay of patients admitted with heart failure”. In: *Journal of hospital medicine* 11.9 (2016), pp. 642–645.

[31] Venkataraghavan Ramamoorthy et al. “Prevalence and Inpatient Hospital Outcomes of Malignancy-Related Ascites in the United States”. In: *American Journal of Hospice and Palliative Medicine®* 38.1 (2021), pp. 47–53.

[32] Monakshi Sawhney et al. “The impact of patient navigation on length of hospital stay and satisfaction in patients undergoing primary hip or knee arthroplasty”. In: *International Journal of Orthopaedic and Trauma Nursing* (2020), p. 100799.

[33] Yan Wu et al. “Does prolonged emergency department length of stay (EDLOS) affect the outcomes of acute ischemic stroke patients?” In: *The American Journal of Emergency Medicine* (2020).

[34] Sangiwe Moyo et al. “Application of machine learning models in predicting length of stay among healthcare workers in underserved communities in South Africa”. In: *Human resources for health* 16.1 (2018), pp. 1–9.

[35] Susanne Weber, Martin Wolkewitz, and COMBACTE-MAGNET Consortium. "Accounting for length of hospital stay in regression models in clinical epidemiology". In: *Statistica Neerlandica* 74.1 (2020), pp. 24–37.

[36] Ryan Healy et al. "Assessment of patient ambulation profiles to predict hospital readmission, discharge location, and length of stay in a cardiac surgery progressive care unit". In: *JAMA network open* 3.3 (2020), e201074–e201074.

[37] Mohamed D Ray-Zack et al. "Hospital length of stay following radical cystectomy for muscle-invasive bladder cancer: Development and validation of a population-based prediction model". In: *Urologic Oncology: Seminars and Original Investigations*. Vol. 37. 11. Elsevier. 2019, pp. 837–843.

[38] Peta MA Alexander et al. "Accurate prediction of congenital heart surgical length of stay incorporating a procedure-based categorical variable". In: *Pediatric Critical Care Medicine— Society of Critical Care Medicine* 19.10 (2018), pp. 949–956.

[39] Carmelo Dominici et al. "A nomogram for predicting long length of stay in the intensive care unit in patients undergoing CABG: Results from the multicenter E-CABG registry". In: *Journal of Cardiothoracic and Vascular Anesthesia* 34.11 (2020), pp. 2951–2961.

[40] Hironori Uematsu et al. "Prediction model for prolonged length of stay in patients with community-acquired pneumonia based on Japanese administrative data". In: *Respiratory Investigation* 59.2 (2021), pp. 194–203.

[41] Rachda Naila Mekhaldi et al. "Using Machine Learning Models to Predict the Length of Stay in a Hospital Setting". In: *World Conference on Information Systems and Technologies*. Springer. 2020, pp. 202–211.

[42] Dhanjeet Singh, Vishal Kumar, and Robin G Qiu. "Patientsâ Disease Risk Predictive Modeling using MIMIC Data". In: *Procedia Computer Science* 168 (2020), pp. 112–117.

[43] Shubham Sinha, Shalini Goel, et al. "Research on Data Science Ensembles for Covid-19 Detection and Length of Stay Prediction". In: *2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*. IEEE. 2021, pp. 499–503.

[44] Whitney E Muhlestein et al. "Predicting inpatient length of stay after brain tumor surgery: Developing machine learning ensembles to improve predictive performance". In: *Neurosurgery* 85.3 (2019), pp. 384–393.

[45] Shilpa Balan, Tanvi Gawade, and Aakanksha Tasgaonkar. "A Machine Learning Approach for Prediction of Length of Stay for the Kidâs Inpatient Database". In: *2019 IEEE International Conference on Big Data (Big Data)*. IEEE. 2019, pp. 5980–5982.

[46] McGraw-Hill Concise Dictionary of Modern Medicine. *length of stay*. 2002. URL: https://medical-dictionary.thefreedictionary.com/length+of+stay (visited on 03/05/2021).

[47] Ardeshir Mansouri, Mohammadreza Noei, and Mohammad Saniee Abadeh. "Predicting Hospital Length of Stay of Neonates Admitted to the NICU Using Data Mining Techniques". In: *2020 10th International Conference on Computer and Knowledge Engineering (ICCKE)*. IEEE. 2020, pp. 629–635.

[48] Roberto Williams Batista and Reinaldo Sanchez-Arias. "A Methodology for Estimating Hospital Intensive Care Unit Length of Stay Using Novel Machine Learning Tools". In: *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE. 2020, pp. 827–832.

[49] Belal Alsinglawi et al. "Predicting Length of Stay for Cardiovascular Hospitalizations in the Intensive Care Unit: Machine Learning Approach". In: *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE. 2020, pp. 5442–5445.

[50] Belal Alsinglawi et al. "Benchmarking Predictive Models in Electronic Health Records: Sepsis Length of Stay Prediction". In: *International Conference on Advanced Information Networking and Applications*. Springer. 2020, pp. 258–267.

[51] Mohammadreza Nemati, Jamal Ansary, and Nazafarin Nemati. "Machine-learning approaches in COVID-19 survival analysis and discharge-time likelihood prediction using clinical data". In: *Patterns* 1.5 (2020), p. 100074.

[52] Peng Zhang et al. "Building cloud-based healthcare data mining services". In: *2016 IEEE international conference on services computing (SCC)*. IEEE. 2016, pp. 459–466.

[53] Jaret M Karnuta et al. "Bundled care for hip fractures: a machine-learning approach to an untenable patient-specific payment model". In: *Journal of orthopaedic trauma* 33.7 (2019), pp. 324–330.

[54] Ali Jalali et al. "Deep learning for improved risk prediction in surgical outcomes". In: *Scientific reports* 10.1 (2020), pp. 1–13.

[55] Xin Ma et al. "Length of stay prediction for ICU patients using individualized single classification algorithm". In: *Computer methods and programs in biomedicine* 186 (2020), p. 105224.

[56] Gaspard Harerimana, Jong Wook Kim, and Beakcheol Jang. "A deep attention model to forecast the Length Of Stay and the in-hospital mortality right on admission from ICD codes and demographic data". In: *Journal of Biomedical Informatics* 118 (2021), p. 103778.

[57] Tahmina Zebin, Shahadate Rezvy, and Thierry J Chaussalet. "A deep learning approach for length of stay prediction in clinical settings from medical records". In: *2019 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*. IEEE. 2019, pp. 1–5.

[58] Sanjay Purushotham et al. "Benchmarking deep learning models on large healthcare datasets". In: *Journal of biomedical informatics* 83 (2018), pp. 112–134.

[59] Baran Balkan and Vignesh Subbian. "Evaluating ICU clinical severity scoring systems and machine learning applications: APACHE IV/IVa case study". In: *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE. 2018, pp. 4073–4076.

[60] Mani Sotoodeh and Joyce C Ho. "Improving length of stay prediction using a hidden Markov model". In: *AMIA Summits on Translational Science Proceedings* 2019 (2019), p. 425.

[61] Lama Alturki et al. "Predictors of Readmissions and Length of Stay for Diabetes Related Patients". In: *2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA)*. IEEE. 2019, pp. 1–8.

[62] Tao Wang et al. "Prediction of hospital length of stay to achieve flexible healthcare in the field of Internet of Vehicles". In: *Transactions on emerging telecommunications technologies* (2019), e3854.

[63] Julia Elrod et al. "Using Artificial Intelligence to Obtain More Evidence? Prediction of Length of Hospitalization in Pediatric Burn Patients". In: *Frontiers in Pediatrics* 8 (2020).

[64] Mahmoud Elbattah and Owen Molloy. "Using machine learning to predict length of stay and discharge destination for hip-fracture patients". In: *Proceedings of SAI Intelligent Systems Conference*. Springer. 2016, pp. 207–217.

[65] Thanos Gentimis et al. "Predicting hospital length of stay using neural networks on mimic iii data". In: *2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech)*. IEEE. 2017, pp. 1194–1201.

[66] Ahmed Belderrar and Abdeldjebar Hazzab. "Hierarchical genetic algorithm and fuzzy radial basis function networks for factors influencing hospital length of stay outliers". In: *Healthcare informatics research* 23.3 (2017), pp. 226–232.

[67] Sergio M Navarro et al. "Machine learning and primary total knee arthroplasty: patient forecasting for a patient-specific payment model". In: *The Journal of arthroplasty* 33.12 (2018), pp. 3617–3623.

[68] Prem N Ramkumar et al. "Development and validation of a machine learning algorithm after primary total hip arthroplasty: applications to length of stay and payment models". In: *The Journal of arthroplasty* 34.4 (2019), pp. 632–637.

[69] Seyedmostafa Sheikhalishahi, Vevake Balaraman, and Venet Osmani. "Benchmarking machine learning models on multi-centre eICU critical care dataset". In: *PloS one* 15.7 (2020), e0235424.

[70] Li Huang et al. "Patient clustering improves efficiency of federated machine learning to predict mortality and hospital stay time using distributed electronic medical records". In: *Journal of biomedical informatics* 99 (2019), p. 103291.

[71] Emma Rocheteau et al. "Predicting Patient Outcomes with Graph Representation Learning". In: *arXiv preprint arXiv:2101.03940* (2021).

[72] Alistair EW Johnson et al. "MIMIC-III, a freely accessible critical care database". In: *Scientific data* 3.1 (2016), pp. 1–9.

[73] Swaraj Khadanga et al. "Using clinical notes with time series data for ICU management". In: *arXiv preprint arXiv:1909.09702* (2019).

[74] Mahsa Rouzbahman, Aleksandra Jovicic, and Mark Chignell. "Can cluster-boosted regression improve prediction of death and length of stay in the ICU?" In: *IEEE journal of biomedical and health informatics* 21.3 (2016), pp. 851–858.

[75] Huan Song et al. "Attend and diagnose: Clinical time series analysis using attention models". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 32. 1. 2018.

[76] Hrayr Harutyunyan et al. "Multitask learning and benchmarking with clinical time series data". In: *Scientific data* 6.1 (2019), pp. 1–18.

[77] Junyi Gao et al. "Dr. Agent: Clinical predictive model via mimicked second opinions". In: *Journal of the American Medical Informatics Association* 27.7 (2020), pp. 1084–1091.

[78] Alistair Johnson et al. "Mimic-iv (version 1.0)". In: *PhysioNet* (2020).

[79] Tom J Pollard et al. "The eICU Collaborative Research Database, a freely available multi-center database for critical care research". In: *Scientific data* 5.1 (2018), pp. 1–13.

[80] Alistair Johnson et al. *MIMIC-IV-ED (version 1.0)*. 2021. DOI: 10.13026/77Z6-9W59. URL: https://physionet.org/content/mimic-iv-ed/1.0/.

[81] Alistair EW Johnson et al. "MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports". In: *Scientific data* 6.1 (2019), pp. 1–8.

[82] Yanbo Xu et al. "Raim: Recurrent attentive and intensive model of multimodal patient monitoring data". In: *Proceedings of the 24th ACM SIGKDD international conference on Knowledge Discovery & Data Mining*. 2018, pp. 2565–2573.

[83] Jen J Gong et al. "Predicting clinical outcomes across changing electronic health record systems". In: *proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2017, pp. 1497–1505.

[84] Brett K Beaulieu-Jones et al. "Characterizing and managing missing structured data in electronic health records: data analysis". In: *JMIR medical informatics* 6.1 (2018), e8960.

[85] Philip B Weerakody et al. "A review of irregular time series data handling with gated recurrent neural networks". In: *Neurocomputing* (2021).

[86] Sujin Kim, Woojae Kim, and Rae Woong Park. "A comparison of intensive care unit mortality prediction models through the use of data mining techniques". In: *Healthcare informatics research* 17.4 (2011), p. 232.

[87] Patrick J Thoral et al. "Sharing ICU Patient Data Responsibly Under the Society of Critical Care Medicine/European Society of Intensive Care Medicine Joint Data Science Collaboration: The Amsterdam University Medical Centers Database (AmsterdamUMCdb) Example". In: *Critical care medicine* (2021).

[88] "Attribute-Value Learning". In: *Encyclopedia of Machine Learning*. Ed. by Claude Sammut and Geoffrey I. Webb. Boston, MA: Springer US, 2010, pp. 53–53. ISBN: 978-0-387-30164-8. DOI: 10.1007/978-0-387-30164-8_43. URL: https://doi.org/10.1007/978-0-387-30164-8_43.

[89] Gareth James et al. *An introduction to statistical learning*. Vol. 112. Springer, 2013.

[90] Jan Salomon Cramer. "The origins of logistic regression". In: (2002).

[91] Stephan Dreiseitl and Lucila Ohno-Machado. "Logistic regression and artificial neural network classification models: a methodology review". In: *Journal of biomedical informatics* 35.5-6 (2002), pp. 352–359.

[92] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[93] Vili Podgorelec et al. "Decision trees: an overview and their use in medicine". In: *Journal of medical systems* 26.5 (2002), pp. 445–463.

[94] Tianqi Chen and Carlos Guestrin. "Xgboost: A scalable tree boosting system". In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 2016, pp. 785–794.

[95] Ameet V Joshi. *Machine learning and artificial intelligence*. Springer, 2020.

[96] Zachary C Lipton, John Berkowitz, and Charles Elkan. "A critical review of recurrent neural networks for sequence learning". In: *arXiv preprint arXiv:1506.00019* (2015).

[97] Hassan Ismail Fawaz et al. "Deep learning for time series classification: a review". In: *Data Mining and Knowledge Discovery* 33.4 (2019), pp. 917–963.

[98] Hansika Hewamalage, Christoph Bergmeir, and Kasun Bandara. "Recurrent neural networks for time series forecasting: Current status and future directions". In: *International Journal of Forecasting* 37.1 (2021), pp. 388–427.

[99] Sepp Hochreiter and Jürgen Schmidhuber. "Long short-term memory". In: *Neural computation* 9.8 (1997), pp. 1735–1780.

[100] Sajad Darabi et al. "Taper: Time-aware patient ehr representation". In: *IEEE journal of biomedical and health informatics* 24.11 (2020), pp. 3268–3275.

[101] Inci M Baytas et al. "Patient subtyping via time-aware LSTM networks". In: *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*. 2017, pp. 65–74.

[102] Kyunghyun Cho et al. “Learning phrase representations using RNN encoder-decoder for statistical machine translation”. In: *arXiv preprint arXiv:1406.1078* (2014).

[103] Fenglong Ma et al. “Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks”. In: *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*. 2017, pp. 1903–1911.

[104] Ashish Vaswani et al. “Attention is All you Need”. In: *NIPS*. 2017.

[105] Keiron O’Shea and Ryan Nash. “An introduction to convolutional neural networks”. In: *arXiv preprint arXiv:1511.08458* (2015).

[106] Neena Aloysius and M Geetha. “A review on deep convolutional neural networks”. In: *2017 International Conference on Communication and Signal Processing (ICCSP)*. IEEE. 2017, pp. 0588–0592.

[107] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. “Deep learning”. In: *nature* 521.7553 (2015), pp. 436–444.

[108] Serkan Kiranyaz et al. “1-d convolutional neural networks for signal processing applications”. In: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2019, pp. 8360–8364.

[109] Rein Houthooft et al. “Predictive modelling of survival and length of stay in critically ill patients using sequential organ failure scores”. In: *Artificial intelligence in medicine* 63.3 (2015), pp. 191–207.

[110] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. “Axiomatic attribution for deep networks”. In: *International Conference on Machine Learning*. PMLR. 2017, pp. 3319–3328.

[111] Faiza Khan Khattak et al. “A survey of word embeddings for clinical text”. In: *Journal of Biomedical Informatics: X* 4 (2019), p. 100057.

[112] Scott M Lundberg and Su-In Lee. “A unified approach to interpreting model predictions”. In: *Proceedings of the 31st international conference on neural information processing systems*. 2017, pp. 4768–4777.

[113] Paula Branco, Luís Torgo, and Rita P Ribeiro. “A survey of predictive modeling on imbalanced domains”. In: *ACM Computing Surveys (CSUR)* 49.2 (2016), pp. 1–50.