Heliyon 10 (2024) e26772

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Multimodal risk prediction with physiological signals, medical images and clinical notes

Yuanlong Wang [a], Changchang Yin [a,b], Ping Zhang [a,b,*]

[a] Department of Computer Science and Engineering, The Ohio State University, Columbus, OH 43210, USA
[b] Department of Biomedical Informatics, The Ohio State University, Columbus, OH 43210, USA

ARTICLE INFO

Keywords:
Electronic health records
Multimodal deep learning
Risk prediction
Data fusion

ABSTRACT

The broad adoption of electronic health record (EHR) systems brings us a tremendous amount of clinical data and thus provides opportunities to conduct data-based healthcare research to solve various clinical problems in the medical domain. Machine learning and deep learning methods are widely used in the medical informatics and healthcare domain due to their power to mine insights from raw data. When adapting deep learning models for EHR data, it is essential to consider its heterogeneous nature: EHR contains patient records from various sources including medical tests (e.g. blood test, microbiology test), medical imaging, diagnosis, medications, procedures, clinical notes, etc. Those modalities together provide a holistic view of patient health status and complement each other. Therefore, combining data from multiple modalities that are intrinsically different is challenging but intuitively promising in deep learning for EHR. To assess the expectations of multimodal data, we introduce a comprehensive fusion framework designed to integrate temporal variables, medical images, and clinical notes in EHR for enhanced performance in clinical risk prediction. Early, joint, and late fusion strategies are employed to combine data from various modalities effectively. We test the model with three predictive tasks: in-hospital mortality, long length of stay, and 30-day readmission. Experimental results show that multimodal models outperform uni-modal models in the tasks involved. Additionally, by training models with different input modality combinations, we calculate the Shapley value for each modality to quantify their contribution to multimodal performance. It is shown that temporal variables tend to be more helpful than CXR images and clinical notes in the three explored predictive tasks.

1. Introduction

Electronic Health Records (EHR) are longitudinal electronic records that contain comprehensive information about a patient's health, including structured data like demographics, vital signs, and laboratory test results, as well as unstructured data such as clinical notes and reports. It is used for effectively and efficiently organizing health records [1] and has been widespread nowadays. The United States healthcare system, for example, serves more than 30 million patients each year. Over the seven years between 2008 and 2015, the adoption rate of at least a basic EHR system by non-federal acute care hospitals increased significantly from 9.4% to 83.8% [2]. As of 2021, 78% of office-based physicians and 96% non-federal acute care hospitals adopted a certified EHR [3]. Due

* Corresponding author.
E-mail address: zhang.10631@osu.edu (P. Zhang).

https://doi.org/10.1016/j.heliyon.2024.e26772
Received 6 October 2023; Received in revised form 17 February 2024; Accepted 20 February 2024

Available online 28 February 2024

to the prevailing usage, the EHR databases encapsulate information from a tremendous population, thus presenting an exceptional opportunity for healthcare researchers to carry out data-driven studies to improve human well-being [4].

Machine learning and deep learning techniques have gained popularity in the healthcare industry due to recent advances and successes [10–12]. They hold great promise in deriving meaningful insights from EHR, which can aid in accurately predicting clinical outcomes, such as mortality [13] and readmission [13,14]. Predicting these outcomes can help in the early detection of patient physiological deterioration [15], which facilitates nursing workflow. Therefore, numerous research studies have utilized deep learning techniques to develop predictive models based on EHR. Typically, vital signs, lab test results, historical diagnosis, and medication information are inspected in these models. However, there is more unstructured information in other modalities available during patient admissions, such as clinical notes and radiography outputs, which could be informative for the predictive tasks. Hence, it is an intuitive idea to fully utilize those complementary data from different modalities during the prediction process to improve model performance.

In this study, we concentrate on combining multimodal data (i.e., clinical time series, chest X-ray radiography [CXR], and radiology notes) produced during patient admissions with a general fusion framework to enhance performance in predicting in-hospital mortality, long length of stay, and 30-day readmission. We introduce three different fusion strategies named early fusion, joint fusion, and late fusion to integrate these heterogeneous data modalities. The model is trained and tested with the MIMIC-MM dataset, which is composed by joining MIMIC-IV, MIMIC-CXR, and MIMIC-IV-Note datasets [5–8]. Additionally, we use the Shapley value [9] to figure out the contribution of each modality during the training process in our predictive tasks. Our experiment results show the superiority of multimodal predictive models.

To summarize, the contributions of our work are:

- We propose a multimodal fusion framework with three fusion strategies (early, joint, and late) to combine clinical time series (e.g. vital signs, lab tests) with CXR images and radiology notes in EHR.
- We conduct experiments on real-world datasets and the experimental results in three tasks (i.e. in-hospital mortality, long length of stay, and 30-day readmission) show that the multimodal models outperform the uni-modal models.
- We adopt the Shapley value to estimate the contribution of each modality and the results show that all modalities are helpful for risk predictions, which further demonstrates the feasibility and effectiveness of the proposed fusion strategies.

2. Related work

Medical datasets are vast collections of patient health records from hospitals, which typically encompass various aspects of a patient's health status, such as demographic information, lab tests, vital signs, medical images, diagnosis codes, notes, treatment and medication history, and discharge reports. In the medical domain, researchers have increasingly utilized machine learning strategies in a variety of medical tasks, such as medical predictive modeling, medical recommendations, disease diagnosis, and medical outcome prediction.

Works on EHR time series variables. There are plenty of attempts to leverage time series in electronic health records (EHR) for predictive modeling tasks. RETAIN [10] applied reversed time attention produced by RNN to generate visit level and variable level attention scores for embedding vectors of clinical time series. It takes into account diagnosis, medication, and procedure events to generate input vectors. Similarly, Dipole [16] use a bidirectional RNN to combine multi-visit embeddings. Med2Vec [11] learned visit-level representation from EHR by mining visit sequence information and medical codes co-occurrence information. The representation from Med2Vec is tested by predicting future medical codes and the Clinical Risk Groups (CRG) level with it. Med-BERT [12], BEHRT [17], and G-BERT [18] utilize a BERT-based framework for EHR feature extraction and are employed in diagnosis code or medication prediction tasks. G-BERT also takes into consideration the hierarchical medical ontology structure of the ICD-9 code to enhance the embedding. Ashfaq et al. [14] leveraged LSTM on top of learned EHR embeddings to predict 30-day readmission.

Works on multimodal data input. Medical datasets exhibit multimodal characteristics, with different types of data such as lab tests and vital signs as time-series variables, medical images, and clinical notes as unstructured text. It is natural and promising to take advantage of complementary information from heterogeneous data [19]. Zhang et al. [13] integrated time series variables with unstructured clinical notes in MIMIC-III to perform predictive modeling tasks, using LSTM and CNN for sequential feature extraction. Golovanevsky et al. [20] incorporated clinical test scores, genetic information (SNPs), and MRI scan images for Alzheimer's disease diagnosis. They adapted cross-modal attention and self-attention modules to capture intra- and inter-modality correlation. Huang et al. [21] utilized Electronic Medical Record (EMR) and CT scan images to detect pulmonary embolism with three fusion methods and found that late fusion modal outperformed others. Yao et al. [22] concatenated selected clinical features with 3D CT image features from CNN for pulmonary venous obstruction (PVO) prediction. They also generated a saliency map to find out which area the model focuses on. The authors claimed that the multimodal model has more concentrated activation in the pulmonary area. Yan et al. [23] conducted breast cancer classification by combining pathological images and 29 selected features. They concatenated hidden states from multiple CNN inner layers as the image feature, applied a denoising autoencoder to obtain EMR features, and concatenated features from images and EMR for classification. Nie et al. [24] combined multi-channel medical images, demographical information, and tumor-related features for short overall survival (OS) time prediction. Soenksen et al. [25] proposed an early fusion model experiments with tabular data, time series data, text notes, and chest X-ray on Chest pathology diagnosis, Length-of-Stay prediction, and 48-hour mortality prediction.

Previous research suggests that leveraging heterogeneous data holds great promise in improving performance on downstream tasks, with early fusion being the most common modality fusion method, where different features from multimodal inputs are

directly concatenated as aggregated features for downstream tasks. Besides, joint fusion and late fusion strategies are also present in the field and worth exploring. Therefore, in this study, we conducted experiments on three representative risk prediction tasks [13] with three modality fusion styles: early fusion, joint fusion, and late fusion. The detailed definition of the tasks and fusion strategies is presented in the following sections. Furthermore, we aimed to quantify the contribution of each modality in each task. Inspired by the explainable AI (XAI) works on feature importance explanation and model explanation [26–28], we calculated the Shapley value for the three modalities to show their contribution.

3. Methods

In this section, we introduce our problem setting formally. Firstly, we describe the multimodal data we utilized as model input, which consists of static features, time series variables, CXR images, and radiology notes. Next, we introduce the architecture of our neural network, which consists of data embedding, feature extraction, modal fusion, and classification head.

3.1. Multimodal data description

3.1.1. Patient demographics

Patient demographics provide basic background information for the patient including age, gender, etc. We utilize age, gender, ethnicity, marital status, language, and insurance condition to provide the initial information of the patient to the predictive model. To enhance model robustness, we binned the patient age by 10 years. Therefore, all the variables considered for demographic information are categorical.

Note that demographic information is included in the clinical time series data in our experiment as they are all tabular in the dataset.

3.1.2. Time series data

From the tabular patient records in the EHR database, we take three types of clinical time series events as input:

- **Chart events** refer to charted items that occurred during the patient's stay in the ICU. This may include vital signs such as Heart Rate and other additional information relevant to their health status like O2 saturation pulseoxymetry, and Glasgow Coma Scale [GCS].
- **Lab events** refer to laboratory measurements made for the patient. For example, Glucose in blood.
- **Procedure events** refer to procedures documented during the ICU stay such as Ventilation.

Following the existing work [25], we focus on a subset of variables. For chart events, we used 6 numeric vital signs and 3 categorical features from the original feature list. For lab events, we focus on 22 lab test items. For procedure events, we take 10 specific operations. The full variable list can be found in Table 1.

Chart events and lab events are clinical variables, every event corresponds to either a numeric measurement, such as for blood glucose concentration, or a categorical measurement like GCS score, so the meaning of their values is clear. However, procedure events are clinical operations or interventions that do not intuitively have value. Therefore, a categorical procedure event means that this procedure is an instant movement that finishes soon, and its value is binary, which denotes if it is conducted at a certain timestamp. For example, the "Chest Tube Removed" event denotes removing the chest tube for the patient at a certain time. On the other hand, a numeric procedure event denotes a continuous procedure that holds for a certain period and the numeric value denotes its duration. The embedding process is described formally in section 3.1.5.

3.1.3. CXR data

CXR data contain the chest X-ray images obtained during patient admission. A patient may have several medical radiology studies during admission and take multiple radiographs in one study. We assume that this image series contains information about patient health status progression. Therefore, it is more reasonable to regard CXR data as an image time series instead of solely taking the most recent image. Since 90% of the CXR images in our setting are anteroposterior, we just use the anteroposterior images as our CXR data branch for simplicity.

3.1.4. Clinical notes

During patient admission, there are various clinical notes about their medical studies, diagnosis, discharge report, etc. For example, the MIMIC-IV-Note Dataset contains radiology notes and discharge notes during patient admission. The deidentified notes and the note date are provided in the unstructured free text. Since discharge notes may contain death information and diagnosis results, we just take the radiology notes from the dataset to avoid possible overfitting and shortcuts.

Radiology notes contain note records for multiple imaging modalities: X-ray, computed tomography, magnetic resonance imaging, ultrasound, etc. Therefore, it is not only a supplement to the CXR modality but a complement to patient admission.

3.1.5. Patient record configuration

A patient X in the dataset is uniquely identified by a patient ID and an admission ID. We can construct and formalize the multimodal patient record by joining clinical time series, CXR records, and Note records on the two ID fields. The record of the

Table 1
Data statistics of MIMIC-IV. For instant procedure events (i.e. Intubation, Bronchoscopy, EEG, and Chest Tube Removed), the number of event occurrences and occurrence rate are reported. For other procedure events (continuous procedures), the mean and deviation of their duration in hours are reported. For patient gender, we also reported the number of male patients and its ratio in the group. Other (numeric) variable statistics are presented in the form of mean (standard deviation).

		In-hospital mortality		Long length of stay		30-day readmission	
		yes	no	yes	no	yes	no
# of admissions (intersection)		3086 (1521)	28002 (10115)	14931 (6137)	11041 (4058)	2429 (494)	28659 (11142)
# of CXR images (intersection)		3448 (3266)	19834 (18938)	13789 (13152)	9493 (6794)	1049 (1005)	22233 (21199)
# of notes (intersection)		7977 (5548)	55312 (31367)	37784 (22675)	25505 (11008)	4123 (1594)	59166 (35321)
demographic	Age (SD)	69.0 (15.6)	62.6 (16.9)	63.3 (16.2)	64.2 (17.2)	65.1 (14.6)	63.3 (17.0)
	Gender (male # and ratio)	841 (55%)	5547 (55%)	3376 (55%)	2192 (55%)	273 (55%)	6115 (55%)
chart events	Heart Rate	91.6 (20.8)	86.3 (18.6)	88.2 (19.1)	85.2 (18.2)	87.1 (18.5)	87.0 (19.1)
	NIBP systolic	111.7 (13.4)	114.3 (13.5)	113.6 (13.5)	114.5 (13.4)	113.1 (13.3)	114.0 (13.5)
	NIBP diastolic	70.5 (7.9)	71.2 (8.1)	71.0 (8.1)	71.1 (8.1)	70.5 (8.2)	71.1 (8.1)
	NIBP mean	77.6 (13.6)	79.7 (13.9)	79.1 (13.9)	79.7 (13.7)	78.4 (14.0)	79.5 (13.8)
	Respiratory Rate	21.4 (6.2)	19.6 (5.5)	20.0 (5.7)	19.5 (5.4)	20.1 (5.8)	19.8 (5.7)
	SpO2	97.2 (2.7)	97.0 (2.5)	97.2 (2.5)	96.9 (2.5)	97.2 (2.5)	97.1 (2.5)
	GCS Verbal Response	2.0 (1.6)	3.2 (1.9)	2.7 (1.9)	3.6 (1.8)	3.0 (1.9)	3.0 (1.9)
	GCS Eye Opening	2.3 (1.3)	3.2 (1.1)	2.9 (1.2)	3.4 (1.0)	3.1 (1.1)	3.1 (1.2)
	GCS Motor Response	4.2 (1.9)	5.4 (1.3)	5.1 (1.5)	5.5 (1.3)	5.2 (1.5)	5.2 (1.5)
lab events	Glucose	156.9 (88.9)	145.1 (67.9)	147.4 (68.6)	143.5 (68.7)	149.7 (75.2)	146.7 (71.4)
	Potassium	4.2 (0.8)	4.1 (0.7)	4.1 (0.7)	4.1 (0.7)	4.2 (0.7)	4.1 (0.7)
	Sodium	138.4 (7.0)	137.9 (5.7)	137.9 (6.1)	137.9 (5.4)	137.8 (5.6)	138.0 (5.9)
	Chloride	104.0 (8.4)	103.7 (6.9)	103.7 (7.4)	103.8 (6.6)	103.3 (6.8)	103.8 (7.2)
	Creatinine	1.9 (1.7)	1.6 (1.7)	1.7 (1.8)	1.4 (1.4)	1.7 (1.5)	1.6 (1.7)
	Urea Nitrogen	36.7 (26.7)	28.3 (22.5)	31.2 (24.8)	26.8 (21.0)	34.3 (24.9)	29.4 (23.3)
	Bicarbonate	21.3 (5.7)	23.4 (5.0)	23.0 (5.1)	23.6 (5.0)	23.8 (5.3)	23.1 (5.2)
	Anion Gap	17.3 (5.8)	14.9 (4.1)	15.4 (4.4)	14.8 (4.0)	15.0 (4.2)	15.3 (4.5)
	Hemoglobin	10.0 (2.2)	10.3 (2.1)	10.2 (2.1)	10.4 (2.1)	9.9 (2.1)	10.2 (2.2)
	Hematocrit	31.1 (6.6)	31.3 (6.2)	30.8 (6.2)	31.6 (6.1)	30.2 (6.0)	31.3 (6.3)
	Magnesium	2.1 (0.4)	2.0 (0.5)	2.0 (0.5)	2.0 (0.5)	2.1 (0.5)	2.0 (0.5)
	Platelet Count	183.9 (124.9)	189.2 (107.4)	187.1 (116.3)	189.8 (98.2)	188.5 (117.2)	188.4 (109.8)
	Phosphate	4.1 (1.9)	3.6 (1.4)	3.7 (1.5)	3.5 (1.3)	3.9 (1.5)	3.7 (1.5)
	White Blood Cell	14.5 (13.0)	11.8 (10.6)	12.6 (12.0)	11.5 (9.0)	13.3 (19.5)	12.2 (10.4)
	Calcium, Total	8.3 (1.0)	8.3 (0.8)	8.3 (0.9)	8.4 (0.8)	8.4 (0.7)	8.3 (0.9)
	MCH	30.2 (3.1)	30.0 (2.7)	30.0 (2.8)	30.0 (2.7)	29.7 (2.8)	30.0 (2.7)
	Red Blood Cells	3.4 (0.8)	3.5 (0.7)	3.4 (0.7)	3.5 (0.7)	3.4 (0.7)	3.5 (0.7)
	MCHC	32.3 (1.8)	32.8 (1.6)	32.7 (1.7)	32.8 (1.7)	32.4 (1.6)	32.7 (1.7)
	MCV	93.4 (8.6)	91.5 (7.2)	91.8 (7.5)	91.5 (7.2)	91.5 (7.5)	91.8 (7.4)
	RDW	16.2 (2.7)	15.3 (2.3)	15.6 (2.4)	15.1 (2.3)	16.1 (2.6)	15.4 (2.4)
	Neutrophils	74.9 (21.3)	77.3 (15.5)	77.0 (17.3)	77.6 (14.1)	77.2 (15.3)	76.8 (16.7)
	Vancomycin	17.4 (8.2)	16.3 (8.5)	16.5 (8.3)	16.7 (8.9)	18.3 (8.4)	16.4 (8.4)
procedure events	Foley Catheter	140.2 (141.2)	87.8 (97.1)	126.6 (122.0)	47.9 (34.3)	109.8 (126.1)	94.8 (105.2)
	PICC Line	116.9 (115.5)	85.7 (97.7)	113.7 (117.6)	55.2 (41.3)	94.9 (104.0)	90.9 (101.6)
	Peritoneal Dialysis	114.6 (70.8)	57.0 (95.5)	75.8 (108.2)	31.2 (34.4)	NaN	61.7 (94.3)
	Dialysis - CRRT	38.4 (39.3)	62.0 (49.3)	66.7 (50.2)	55.3 (28.2)	49.7 (24.7)	52.2 (47.5)
	Dialysis Catheter	77.6 (87.0)	103.4 (106.8)	125.7 (114.0)	54.8 (34.8)	132.1 (167.1)	94.5 (97.9)
	Hemodialysis	19.8 (40.4)	15.1 (29.0)	15.9 (35.8)	18.2 (25.8)	9.3 (11.1)	16.0 (31.3)
	Intubation	239 (15.1%)	857 (8.1%)	728 (11.4%)	271 (6.4%)	45 (8.7%)	1051 (9.1%)
	Bronchoscopy	103 (6.2%)	324 (2.9%)	292 (4.3%)	98 (2.2%)	23 (4.3%)	404 (3.3%)
	EEG	123 (7.4%)	347 (3.3%)	279 (4.3%)	144 (3.2%)	13 (2.4%)	457 (3.9%)
	Chest Tube Removed	6 (0.4%)	205 (2.0%)	75 (1.2%)	134 (3.2%)	10 (2.0%)	201 (1.7%)

patient x denoted as R_x is a tuple (D_x, S_x, M_x, N_x). Here $D_x \in \mathbb{R}^{|D|}$ is the demographic information of the patient, and $|D|$ is the number of variables considered for demographic information. Unnecessary subscripts for patient identification are ignored below to avoid confusion.

$S = (c, l, p)$ is the clinical time series of the patient, it consists of three types of events: chart events c, lab tests l, and procedure events p. c and l are two sets of variables, a variable v is a set of timestamped value $v = \{(value_i, t_i) | t_i \in T_v\}$, T_v is the set of observed time for variable v. The variables can be either numeric or categorical. Numeric variables take values in $\mathbb{R}$ while categorical variables take values in a finite set. $p = (o_i, o_c)$ is the union of instant and continuous operations, o_i is a set of instant operations of the form (f, t) which means the operation f takes place at time t; o_c is a set of continuous operations of the form (f, t_s, t_e) which means the operation f starts at t_s and ends at t_e.

$M = \{(m_i, t_i) | t_i \in \mathbb{R}_+\}$ is the CXR record of the patient, it is a set of timestamped CXR images. Here t_i refers to the CXR study time for image m_i. Every m_i is a three dimension tensor $m_i \in \mathbb{R}^{H \times W \times C}$ where H, W, C refer to height, weight, and channel.

$N = \{(n_i, t_i) | t_i \in \mathbb{R}_+\}$ is the Note record of the patient, a set of timestamped radiology notes. t_i denotes the chart time of the patient's notes. n_i is a string of the deidentified clinical notes.

The overview of patient records is shown in Fig. 1. In a nutshell, the patient record is the combination of static demographic information and multivariate time series (i.e. clinical series, CXR image series, radiology note series).

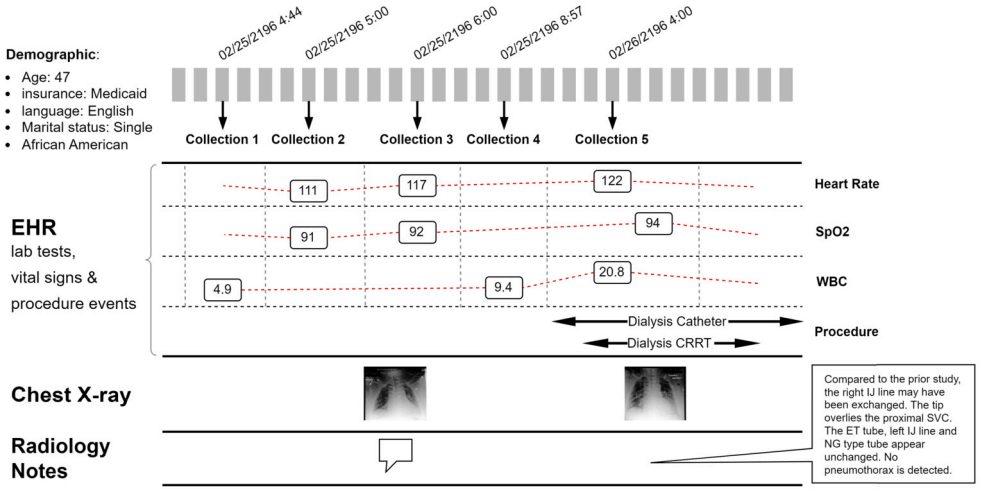

Fig. 1. Patient Record Overview. We use demographical information, multiple temporal signs, chest X-ray images, and clinical radiology notes from the patient admission records and each event has an exact timestamp indicating the time for that event. Continuous events have two timestamps indicating their start or end time.

3.2. Neural network architecture

We introduce the designed architecture of our general predictive model in this subsection. The model can be decomposed into data embedding, uni-modal feature extraction, time series representation, classifier, and modality fusion module. We fuse the modalities with 3 different strategies: early fusion, joint fusion, and late fusion.

3.2.1. Data embedding

The length of stay and number of events varies a lot among patients. The ranges of value of each variable are also different. Therefore, we need to further process the data before feeding them into the model. Additionally, the variable time series are further embedded into more expressive vectors.

Given certain patient demographics and time series (D, S), we first combine and transform them to:

$$S' = \{(d, c_t, l_t, p_t, t) | t \in \mathbb{R}_+\}$$

where $d \in \mathbb{R}^{|D|}$ is the vector for demographics, t is the timestamp from clinical time series of x. $c_t = \{(v_i, val_i) | v_i = val_i \ at \ t\}$ is the set of observed variables from chart events and their values at time t. l_t is the set of lab events and has the same form as c_t. p_t is the set of procedure events that the patients get at time t. The instant procedure conducted at t occurs once in p_t, while the continuous procedure occurs in all p_t that have t between the operation start time and end time.

After the transformation of clinical time series, we use three kinds of embedding: variable embedding, value embedding, and time embedding [29].

Variable embedding encodes what the variable is into a vector, different variables have different embeddings. Given N variables, the variable embedding for the variable v is a linear map from its one hot representation $r_v \in \{0, 1\}^N$ to a vector $e^v = W r_v, W \in \mathbb{R}^{d \times N}$, where d is the embedding size.

Value embedding encodes the value of variables into a vector. For categorical variables, including demographic features, value embedding is a map from its value range set to a real value vector. For numeric variables, we discretize the values into L sub-ranges according to all observed values in the database ensuring that each sub-range contains the same number of samples (equal-frequency binning). Therefore, the discretized values are uniformly distributed in the subranges. Then for a sub-range $1 \le l \le L$, it is embedded into a vector $e^l \in \mathbb{R}^{2k}$ by:

$$e^l_j = \sin(\frac{l \times j}{L \times k}) \qquad e^l_{k+j} = \cos(\frac{l \times j}{L \times k})$$

where $1 \le j \le k$.

Thus, given the event that a variable $v = val$ at time t, we can get variable embedding $e^v \in \mathbb{R}^d$, and value embedding $e^{val} \in \mathbb{R}^{2k}$, where d and $2k$ are predefined embedding sizes and we set $d = 2k$. Then we use a linear function to map the concatenation $[e^v, e^{val}] \in \mathbb{R}^{2d}$ to $e^{var} \in \mathbb{R}^d$ as the embedding of the event that a numeric variable v have value val. This embedding strategy can be employed for all variables of demographic features, chart events, lab events, and procedure events.

Then we try to fuse timestamps into the embedding process. By calculating the relative time from the patient admission, timestamps are converted to real numbers and therefore can be discretized and embedded as $e^t \in \mathbb{R}^d$ with the same value embedding strategy as for numeric variables.

Given a patient X, there can be multiple events (variables) observed at time t. We use adaptive max pooling to extract important information from those embeddings. Recall that for any observed variable and its value, we embed them as $e^{var} \in \mathbb{R}^d$. Hence, the set of observed variables at time t forms a set of embedding $\{e^{var_i} | var_i \text{ observed at } t\}$. Through stacking demographic embedding $E^D \in \mathbb{R}^{|D| \times d}$ and embeddings of all variables, we get an embedding matrix $E_t^{ts} = vstack(e^D, e^{var_i})$, $E_t^{ts} \in \mathbb{R}^{* \times d}$ where $*$ is decided by the number of variables observed. After adaptive max pooling, we conclude the embedding at time t as $e_t^{ts} \in \mathbb{R}^d$. Then we concatenate it with the time embedding e^t and get the final record embedding at time t as $e_t'^{ts} \in \mathbb{R}^{2d}$.

The process for CXR images is simple. The original CXR images are large grayscale images. In order to fit images into our ResNet feature extractor, we resized them to 224 × 224 and duplicated them across 3 input channels. After that, we get images $m \in \mathbb{R}^{3 \times 224 \times 224}$.

For free text notes, typical NLP transformations are applied to convert natural sentences to token lists. All words in the notes are converted to lowercase and tokenized to form a word sequence, punctuations are removed. For example, a sentence like "History of diarrhea and malaise, now with cardiac arrest." becomes a sequence: history, of, diarrhea, and, malaise, now, with, cardiac, arrest. Each word is transformed into an integer through a predefined dictionary, so we get note embedding $n \in \mathbb{N}^L$, here L denotes note length.

After such an embedding process, we got a tensor representation of each modality. Then we use different backbone neural networks to extract feature vectors from these modalities.

3.2.2. Modal feature extraction

After data embedding, we use neural network architecture to extract features from them and produce feature vectors for classification.

Clinical time series features are more relevant to the time dimension. Hence, we use a bidirectional LSTM network for its ability to recall long-term information. As mentioned above, after the embedding procedure, the record at time t can be represented as $e_t'^{ts} \in \mathbb{R}^{2d}$. Thus, for any patient X, we have $E^{ts} \in \mathbb{R}^{T_{ts} \times 2d}$ where T_{ts} is the number of timestamps from clinical time series. We put it into a bidirectional LSTM network. The procedure can be described as follows:

$$\overrightarrow{h_1}, \overrightarrow{h_2}, \cdots, \overrightarrow{h_{T_{ts}}} = \overrightarrow{LSTM}(e_1'^{ts}, e_2'^{ts}, \cdots, e_{T_{ts}}'^{ts})$$
$$\overleftarrow{h_1}, \overleftarrow{h_2}, \cdots, \overleftarrow{h_{T_{ts}}} = \overleftarrow{LSTM}(e_1'^{ts}, e_2'^{ts}, \cdots, e_{T_{ts}}'^{ts})$$
$$h_i = concat[\overrightarrow{h_i}, \overleftarrow{h_i}],\ \forall i \in \{1, 2, \cdots, T_{ts}\}$$

Here $\overrightarrow{LSTM}$ and $\overleftarrow{LSTM}$ are the forward pass and backward pass of bidirectional LSTM, respectively. $\overrightarrow{h_i}, \overleftarrow{h_i} \in \mathbb{R}^d$, $h_i \in \mathbb{R}^{2d}$. After the LSTM layer, we keep the most important information in the series $h_1, h_2, \cdots, h_{T_{ts}}$ by max pooling and take the output as the final feature vector extracted from the clinical time series.

$$\hat{E}^{ts} = maxpooling(h_1, h_2, \cdots, h_{T_{ts}}) \in \mathbb{R}^{2d}$$

We use ResNet for image feature extraction. The original classification head of ResNet is substituted with a Linear layer that generates feature vectors from the output of the convolution layers. After that, an image m_i turns into a vector $e^{m_i} \in \mathbb{R}^{2d}$ For any patient X, we get $E^{cxr} \in \mathbb{R}^{T_{cxr} \times 2d}$ as the features of CXR images at different timestamps, here T_{cxr} is the number of CXR images. After that, we do a weighted sum of T_{cxr} feature vectors according to the time gap between their timestamps and the patient's admission time. Given image features $E^{cxr} = (e^{m_1}, e^{m_2}, \cdots, e^{m_{T_{cxr}}})^T \in \mathbb{R}^{T_{cxr} \times 2d}$ and their time gap from admission $t_1, t_2, \cdots, t_{T_{cxr}}$, their weighted sum over time is defined as:

$$\hat{E}^{cxr} = softmax(t_1, t_2, \cdots, t_{T_{cxr}}) E^{cxr} \in \mathbb{R}^{2d}$$

The weighted sum serves as the final feature vector extracted from CXR records. By softmax function, the latest image feature has the largest weight and the weight decays exponentially as the time gap increases.

For free text notes, we train a Doc2Vec module [30] with radiology notes in the training set to serve as a feature extractor of the free text modality that maps a token sequence n_i to a representation vector $e^{n_i} \in \mathbb{R}^{2d}$. For the patient X, $E^{note} \in \mathbb{R}^{T_{note} \times 2d}$ is produced to serve as the feature vector time series corresponding to the patient note sequence. T_{note} is the number of clinical notes. Given the feature series, we further capture the element correlation and sequence characteristic by an LSTM network [31]. Given $E^{note} = (e^{n_1}, e^{n_2}, \cdots, e^{n_{Tnote}})^T \in \mathbb{R}^{T_{note} \times 2d}$, we fed it through an LSTM network and do max-pooling over all hidden states to generate a single feature vector containing information from the entire sequence:

$$h_1, h_2, \cdots, h_{T_{note}} = LSTM(e^{n_1}, e^{n_2}, \cdots, e^{n_{Tnote}})$$
$$\hat{E}^{note} = maxpooling(h_1, h_2, \cdots, h_{T_{note}}) \in \mathbb{R}^{2d}$$

3.2.3. Classifier

The classifier is built on top of the extracted features to classify them into negative class and positive class. The meaning of the two classes varies according to the predictive task we work on. For example, in in-hospital mortality prediction, negative means the patient was alive at discharge, and positive means the opposite.

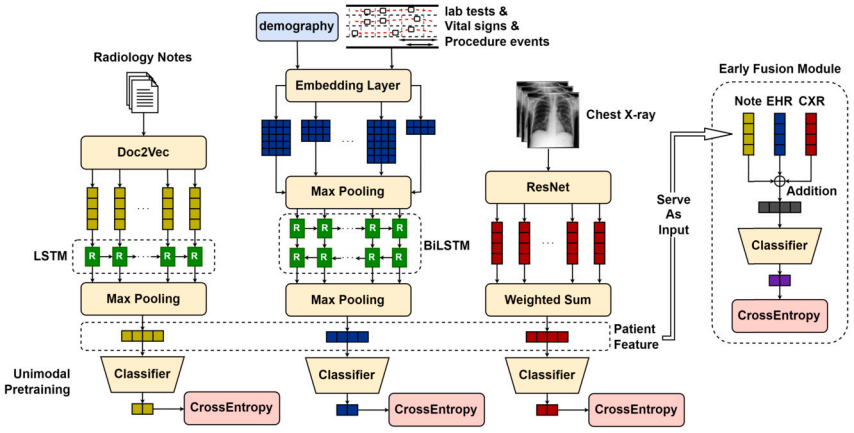

Fig. 2. Early fusion model structure. The feature extractors of each modality are trained in advance on the target tasks. After the convergence of separate training, extractors with the best AUROC score on the testing dataset are fixed to extract patient features from each patient sample. The features are used to train the final classifier.

We employ the linear classifier for all model settings. The linear classifier is simply a fully connected layer of the form:

$$f_{cls}(x) = xW_{cls} + b_{cls}$$

where $W_{cls} \in \mathbb{R}^{2d \times 2}$ is the weight and $b \in \mathbb{R}^2$ is the bias. The input x depends on the fusion method we use. With joint or early fusion, $x = sum[\hat{E}^{ts}, \hat{E}^{cxr}, \hat{E}^{note}]$. With late fusion, the three modalities have their own classifiers respectively and their results are averaged to form the final output, so $x = \hat{E}^m, m \in \{ts, cxr, note\}$. We explain more about the fusion method in section 3.2.4.

The output of the classifier is followed by the softmax function to get the predicted probabilities of each class and the cross entropy is used to measure the classification loss.

3.2.4. Multimodal fusion

Based on the feature vector extracted from the former steps, inspired by Kline et al. [19] and following the definition of [32], we employ three fusion strategies to fuse the event time series feature, the CXR feature and the radiology note feature to generate prediction. The methods are early fusion, joint fusion, and late fusion.

Early fusion joins feature vectors from multiple modalities before feeding them into the classification network. In practice, we directly add up the features from different modalities to form a multimodal feature vector. After that, we feed it into the classifier and get classification results. In this case, the input dimension of the classification layer is the same as single modality feature dimensions, which is $2d$ in our case. For any prediction task, the feature extractor is trained on each modality respectively with their separate classifiers. After pretraining separately, we fix the feature extractor and fuse feature vectors from different modalities to train the multimodal classifier. The process is shown in Fig. 2.

Joint fusion combines the learned features from intermediate layers of different neural networks for different modalities. The difference between joint fusion and early fusion here is that early fusion leverages invariant features pretrained on each modality respectively while joint fusion trains an end-to-end model that propagates gradients to each feature extractor from the multimodal classifier. The network structure is nearly the same as early fusion but the training strategy is different. Direct addition is also used here to construct multimodal feature vectors so the input dimension of the classification layer is also $2d$. The structure of joint fusion is shown in Fig. 3.

Late fusion trains different classifiers for modalities respectively and combines their uni-modal prediction to form a global multimodal prediction. This strategy resembles ensemble learning and is also known as decision-level fusion. There are different styles of assembling predictions, we select averaging in our implementation. The strategy is shown in Fig. 4.

Now we can reach a conclusion about our multimodal prediction model. Just as Fig. 2, 3, 4 shows, the original data undergoes embedding, feature extraction, modal fusion, and classification to generate final predictions. Note that the model is capable of handling missing modalities. For example, if a patient has no CXR record, we can just ignore the CXR feature extractor and the CXR classifier during the fusion stage without changing network architecture and use the other two modalities to generate predictions.

4. Experiment and discussion

4.1. Data description

Medical Information Mart for Intensive Care IV (MIMIC-IV) contains data from hospital stays for patients who were admitted to the Beth Israel Deaconess Medical Center (BIDMC) between 2008 and 2019. We use the 1.0 version of MIMIC-IV for our experiments. MIMIC-IV is divided into five sections: core (information on patient stays), hospital (laboratory and microbiology data), ICU data

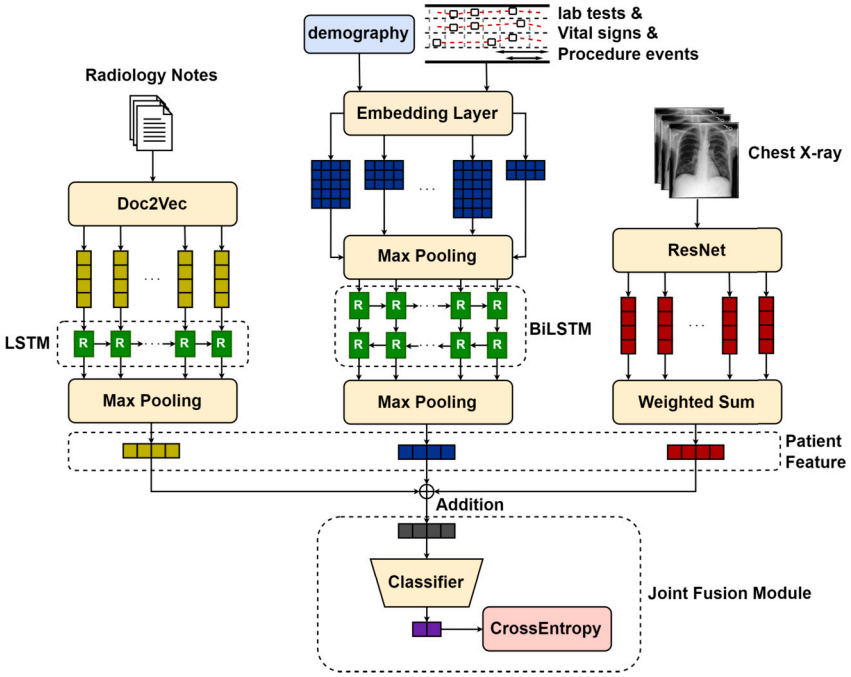

Fig. 3. Joint fusion model structure. The feature extractors are directly connected to the classifier and trained together in one go. The training starts from the random initialization of both feature extractors and classifier, then is trained end-to-end on the multimodal dataset.

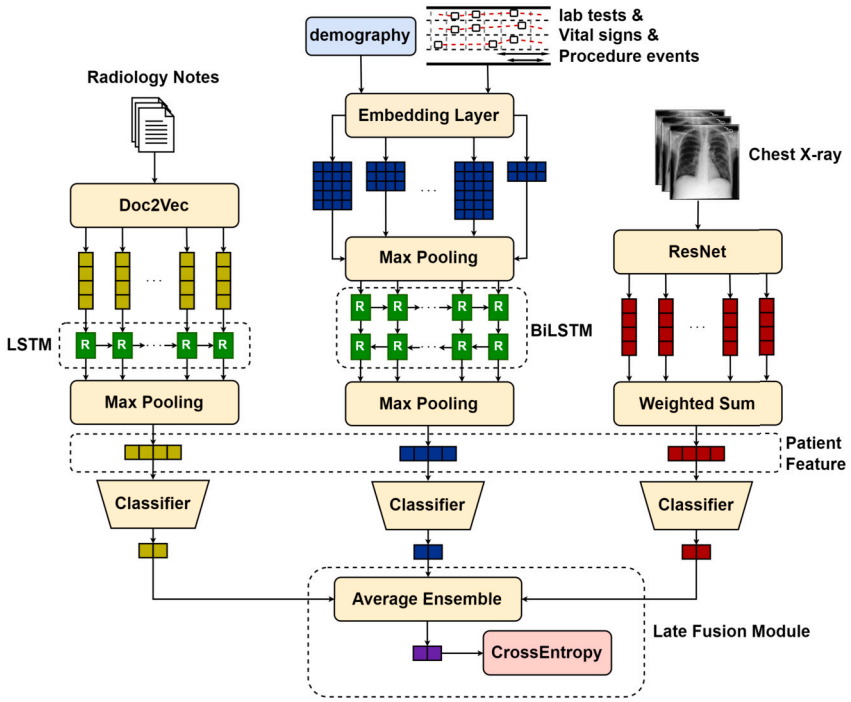

Fig. 4. Late fusion model structure. There are three classifiers attached to each feature extractor for each modality. The predictions of all three modalities are aggregated by calculating the average to produce the final prediction. The late fusion model was also trained in an end-to-end manner.

(details on ICU stays and events), emergency department, and CXR (lookup tables for joining MIMIC-CXR). We utilize clinical time series and patient demographics from MIMIC-IV version 1.0, when we mention clinical time series, it also includes demographics.

MIMIC-CXR is a vast, openly accessible database of patient chest X-rays gathered from the BIDMC emergency department from 2011 to 2017. It includes 227,835 X-ray studies for 64,588 patients. A single study can include several images taken from various view positions, amounting to a total of 377,110 radiographs. Additionally, each study is accompanied by a free-text radiology report created at the time of the study.

MIMIC-IV-Note is an extension of MIMIC-IV on free text clinical notes. Using the same inclusion criteria, MIMIC-IV-Note provides deidentified radiology notes and discharge notes for each patient admission. It contains 331,794 deidentified discharge summaries from 145,915 patients admitted and 2,321,355 deidentified radiology reports for 237,427 patients. All note records in the database can be linked to MIMIC-IV by patient and admission ID numbers.

We use MIMIC-IV-MM [25] to train our model. MIMIC-IV-MM is generated by joining MIMIC-IV, MIMIC-CXR, and MIMIC-IV-Note on the triplet of patient subject ID, hospital admission ID, and ICU-stay ID. Patient records in datasets are aligned to form a universal multimodal patient record in our study.

MIMIC-IV-MM can be seen as an intersection of the three datasets. Therefore, only patients with records in all datasets are included in our study. Even though a patient may lack certain modalities such as CXR images and clinical notes during ICU stay, as stated before, our model is capable of dealing with missing modalities by ignoring them during the fusion phase. However, when training our model, we assume that all modalities are available for every patient.

Moreover, a person can have multiple hospital admissions. However, we treat different admissions of the same person as different patients in our dataset and ignore the latent correlation between them for simplicity.

4.2. Cohort preparation

Based on the MIMIC-IV, MIMIC-CXR, and MIMIC-IV-Note datasets, we evaluated our proposed models on the in-hospital mortality prediction, long length of stay prediction, and readmission prediction based on the record within the first 48 hours after admission. Patients who have all three modalities available are included. In these patients, patients who have no event records within the first 48 hours of their admission are removed. After that, there are 11,636 unique patients left, and the distribution over classes is shown in Table 1. Additional exclusion criteria specific to tasks are introduced in the predictive tasks section below.

4.3. Predictive tasks

We conducted experiments on three prediction problems to test our model. They are all binary classification problems. The detailed definitions of these problems are stated below.

4.3.1. In-hospital mortality prediction

In-hospital mortality is considered a key outcome of interest. The main goal of this task is to forecast if a patient will die during their hospital admission. For any patient, we use events, images, and notes within the first 48 hours from admission as input to the predictive model and generate binary classification indicating whether the patient passes away at discharge. We report the F1 score, the area under the receiver operating characteristic curve (AUROC), the area under the precision-recall curve (AUPRC) of the positive class, precision, recall, and the overall accuracy to measure the performance of the model on this task.

4.3.2. Long length of stay prediction

The length of patient stay refers to the length of time from a patient's admission to discharge. Identifying possible long hospital stays helps in hospital resource management. For simplicity, we formalize the length of stay problem as a binary classification. With observed events, images, and notes in the first 48 hours of admission, the model tries to decide whether the patient will stay in the hospital for more than 7 days [33]. Positive samples are patients that stay for more than 7 days and all other patients are negative samples. The same criteria (AUROC, AUPRC, precision, recall, accuracy) are employed to evaluate model performance. To show a clear margin between methods, we exclude patients that have a stay time shorter than 3 days.

4.3.3. Hospital readmission prediction

It is reported that 13% of the inpatients in the US consume more than half of the hospital resources by readmission [34]. Therefore, it is helpful to have a predictive model to support better readmission prevention and patient satisfaction. We define hospital readmission as unplanned admission within 30 days following the initial discharge [13], which is a binary classification task. Patient data records within the first 48 hours from admission are collected to predict if the patient will be readmitted within 30 days from discharge. The same criteria (AUROC, AUPRC, precision, recall, accuracy) are used to evaluate model performance.

4.4. Implementation details

The model is implemented with PyTorch 1.13.1. All experiment configurations use the weighted cross-entropy loss as the loss function, with 1 for the negative class and 10 for the positive class. In the CXR partial model, the weight for the positive class is 15, "partial" here means the model only utilizes chest X-ray images for training. Models are optimized with the Adam optimizer with 0.001 learning rate until they converge for about 20∼30 epochs. The size of the feature vectors of each modality is set to 512. For

Table 2
Performance result of different fusion methods and modality combination on in-hospital mortality prediction.

	Modality	AUROC	AUPRC	F1	Precision	Recall	Acc
Partial	TimeSeries	0.763	0.448	0.40	0.39	0.40	0.84
	CXR	0.534	0.139	0.14	0.14	0.14	0.77
	Note	0.675	0.255	0.29	0.29	0.30	0.81
Late	T+C	0.799	0.487	0.45	**0.46**	0.44	**0.86**
	T+N	0.813	0.480	**0.48**	**0.46**	**0.49**	**0.86**
	T+C+N	0.823	0.500	0.47	0.45	0.48	**0.86**
Joint	T+C	0.809	0.502	0.45	**0.46**	0.43	**0.86**
	T+N	0.821	**0.508**	0.46	**0.46**	0.46	**0.86**
	T+C+N	**0.825**	**0.508**	0.47	**0.46**	**0.49**	**0.86**
Early	T+C	0.770	0.447	0.41	0.39	0.43	0.84
	T+N	0.788	0.438	0.44	0.43	0.45	0.85
	T+C+N	0.790	0.440	0.44	0.44	0.45	0.85

Table 3
Performance result of different fusion methods and modality combination on the long length of stay prediction.

	Modality	AUROC	AUPRC	F1	Precision	Recall	Acc
Partial	TimeSeries	0.704	0.790	0.71	0.71	0.71	0.65
	CXR	0.565	0.659	0.64	0.65	0.64	0.57
	Note	0.649	0.744	0.69	0.69	0.69	0.62
Late	T+C	0.721	0.804	0.72	0.73	0.71	0.66
	T+N	**0.735**	0.812	0.72	**0.75**	0.70	0.67
	T+C+N	**0.735**	0.812	0.72	0.74	0.71	0.67
Joint	T+C	0.730	0.810	0.72	0.73	0.71	0.66
	T+N	0.730	**0.816**	0.72	0.73	0.71	0.67
	T+C+N	**0.735**	0.812	**0.73**	0.74	0.72	**0.68**
Early	T+C	0.713	0.794	0.71	0.72	0.70	0.66
	T+N	0.714	0.798	0.72	0.73	0.72	0.67
	T+C+N	0.718	0.801	**0.73**	0.73	**0.73**	0.67

evaluation, we use a 0.70-0.15-0.15 train-validation-test split following the previous work [13]. However, we equalized the size of the validation set and the test set to get a better estimation of test performance on the validation set. We split the dataset according to patient ID and admission ID so that there's no overlap between training, validation, and test sets. We use the AUROC on the validation set for early stopping. During every epoch, the model is trained and validated, and the model with the highest AUROC score on the validation set is saved and chosen as the final output model. After that, the result is tested with the saved model on the test set, which is never used during training.

4.5. Results

This section presents the performance results of the proposed models across three tasks: in-hospital mortality prediction, long length of stay prediction, and hospital readmission. For the modality ablation study, we regard clinical time series as the main modality, CXR, and notes as additional ones. Therefore, besides three unimodal experiments, we conduct experiments on TimeSeries + CXR (T + C), TimeSeries + Note (T + N), and TimeSeries + CXR + Note (T + C + N). Unimodal models are denoted as "partial" in the tables.

To validate the performance gap between different model settings, we conducted statistical tests between each pair of model settings. We used a t-test on AUROC across different models and reported the p-values of the comparison.

We also compared the best multimodal models with two baseline models, RETAIN [10] and Dipole [16]. The other EHR models we included in the related works are language modeling-based pretraining models, which we believe are not aligned with our framework.

After showing the performance metrics, we provide the Shapley value as a measurement of the contribution of each modality during training.

4.5.1. Model performance

The results are shown in Table 2, 3, and 4. The best performance in each column is shown in bold, multiple top-1 performance scores are all bolded. It is shown in the table that clinical time series works better among unimodal models for the three tasks, possibly because vital signs and lab tests contain more information about the patient's overall health status, which can be more beneficial for the three tasks. However, the performance can be boosted with additional modalities. The improvement with additional modalities is also consistent over the three fusion strategies.

Table 4
Performance result of different fusion methods and modality combination on hospital readmission prediction.

	Modality	AUROC	AUPRC	F1	Precision	Recall	Acc
Partial	TimeSeries	0.544	0.072	**0.11**	**0.11**	0.10	**0.92**
	CXR	0.525	0.047	0.04	0.03	0.06	0.87
	Note	0.551	0.048	0.01	0.01	0.01	**0.92**
Late	T+C	0.580	0.060	0.10	0.10	0.10	**0.92**
	T+N	0.595	0.065	0.10	0.09	**0.12**	0.91
	T+C+N	0.606	0.066	0.09	0.08	0.10	0.91
Joint	T+C	0.587	**0.073**	0.10	0.09	**0.12**	0.91
	T+N	**0.629**	0.071	0.10	0.09	0.10	**0.92**
	T+C+N	0.513	0.057	0.06	0.05	0.06	0.91
Early	T+C	0.578	0.062	**0.11**	0.10	**0.12**	0.91
	T+N	0.546	0.047	0.01	0.01	0.01	0.91
	T+C+N	0.544	0.047	0.01	0.01	0.01	0.91

Table 5
Performance result of the baseline models and the best multimodal models for each task. Partial denotes the partial model for clinical time series in the tables above. Multimodal denotes the multimodal model with the best AUROC score for the three tasks. LOS: length of stay.

Task	Model	AUROC	AUPRC	F1	Precision	Recall	Acc
Mortality	RETAIN	0.799	0.470	0.45	0.46	0.43	0.86
	Dipole	0.787	0.432	0.42	0.45	0.38	0.86
	Partial	0.763	0.448	0.40	0.39	0.40	0.84
	Multimodal	**0.825**	**0.508**	**0.47**	**0.46**	**0.49**	**0.86**
Long LOS	RETAIN	0.698	0.766	0.71	0.72	0.69	0.65
	Dipole	0.671	0.762	0.69	0.70	0.69	0.63
	Partial	0.704	0.790	0.71	0.71	0.71	0.65
	Multimodal	**0.735**	**0.812**	**0.73**	**0.74**	**0.72**	**0.68**
Readmission	RETAIN	0.574	0.051	0.04	0.03	0.04	0.91
	Dipole	0.560	0.053	0.03	0.03	0.03	0.92
	Partial	0.544	**0.072**	**0.11**	**0.11**	0.10	0.92
	Multimodal	**0.629**	0.071	0.10	0.09	**0.10**	**0.92**

Table 6
P-value of t-test for the in-hospital mortality prediction between different model settings. The model names are in the form "method_modality", where "e" means clinical time series, "c" means chest X-ray, and "n" means radiology notes for modality.

	early_ec	early_ecn	early_en	joint_ec	joint_ecn	joint_en	late_ec	late_ecn	late_en	partial_c	partial_e	partial_n
early_ec	1	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
early_ecn	<0.01	1	0.638	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
early_en	<0.01	0.638	1	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
joint_ec	<0.01	<0.01	<0.01	1	<0.01	0.725	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
joint_ecn	<0.01	<0.01	<0.01	<0.01	1	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
joint_en	<0.01	<0.01	<0.01	0.725	<0.01	1	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
late_ec	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	1	<0.01	<0.01	<0.01	<0.01	<0.01
late_ecn	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	1	<0.01	<0.01	<0.01	<0.01
late_en	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	1	<0.01	<0.01	<0.01
partial_c	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	1	<0.01	<0.01
partial_e	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	1	<0.01
partial_n	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	1

The performance of baseline models is in Table 5. "Partial" denotes the partial model for clinical time series in Table 2, 3, and 4. "Multimodal" denotes the multimodal model with the highest AUROC for each task. The best performance score for each task is bolded. It can be shown from the result that RETAIN [10] and Dipole [16] outperform our clinical time series partial model on mortality and readmission prediction but failed on the long length of stay task. Meanwhile, the multimodal model is significantly better than the unimodal baselines on the three tasks. We believe that this advantage comes from the additional information from multimodality.

4.5.2. Statistical test

We conducted a statistical t-test to validate the comparison between model pairs of fusion methods and partial models for each task. The results are in Table 6, Table 7, and Table 8. The p-value below 0.01 is represented as "<0.01" in the table. We can observe that the AUROC difference between most model pairs is significant with a low p-value.

Table 7
P-value of t-test for the long length of stay prediction between different model settings. The model names are in the form "method_modality", where "e" means clinical time series, "c" means chest X-ray, and "n" means radiology notes for modality.

	early_ec	early_ecn	early_en	joint_ec	joint_ecn	joint_en	late_ec	late_ecn	late_en	partial_c	partial_e	partial_n
early_ec	1	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	0.328	<0.01
early_ecn	<0.01	1	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
early_en	<0.01	<0.01	1	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
joint_ec	<0.01	<0.01	<0.01	1	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
joint_ecn	<0.01	<0.01	<0.01	<0.01	1	0.529	<0.01	<0.01	0.928	<0.01	<0.01	<0.01
joint_en	<0.01	<0.01	<0.01	<0.01	0.529	1	<0.01	<0.01	0.695	<0.01	<0.01	<0.01
late_ec	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	1	<0.01	<0.01	<0.01	<0.01	<0.01
late_ecn	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	1	<0.01	<0.01	<0.01	<0.01
late_en	<0.01	<0.01	<0.01	<0.01	0.928	0.695	<0.01	<0.01	1	<0.01	<0.01	<0.01
partial_c	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	1	<0.01	<0.01
partial_e	0.328	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	1	<0.01
partial_n	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	1

Table 8
P-value of t-test for the hospital readmission prediction between different model settings. The model names are in the form "method_modality", where "e" means clinical time series, "c" means chest X-ray, and "n" means radiology notes for modality.

	early_ec	early_ecn	early_en	joint_ec	joint_ecn	joint_en	late_ec	late_ecn	late_en	partial_c	partial_e	partial_n
early_ec	1	0.012	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
early_ecn	0.012	1	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
early_en	<0.01	<0.01	1	0.261	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
joint_ec	<0.01	<0.01	0.261	1	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
joint_ecn	<0.01	<0.01	<0.01	<0.01	1	<0.01	<0.01	<0.01	<0.01	<0.01	0.025	<0.01
joint_en	<0.01	<0.01	<0.01	<0.01	<0.01	1	<0.01	0.836	<0.01	<0.01	<0.01	<0.01
late_ec	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	1	<0.01	<0.01	<0.01	<0.01	0.139
late_ecn	<0.01	<0.01	<0.01	<0.01	<0.01	0.836	<0.01	1	<0.01	<0.01	<0.01	<0.01
late_en	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	1	<0.01	<0.01	<0.01
partial_c	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	1	<0.01	<0.01
partial_e	<0.01	<0.01	<0.01	<0.01	0.025	<0.01	<0.01	<0.01	<0.01	<0.01	1	<0.01
partial_n	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	0.139	<0.01	<0.01	<0.01	<0.01	1

4.5.3. Shapley value calculation

Shapley value is a concept in cooperative game theory that distributes the total surplus reached by the player coalition to every coalition member according to the player's marginal contribution. The value is constrained by a collection of axioms so that it is the unique solution satisfying the constraints. This concept is also widely used in explainable AI [26–28] to explain the contribution of features and samples, etc.

In our circumstance, our goal is to measure the contribution of the information from each modality to the performance (AUROC) of the trained model during the training process. Let $N = \{t, c, n\}$ be the modality set where t denotes clinical time series, c denotes CXR images, and n denotes radiology notes, we define a function $v : 2^N \to \mathbb{R}$ that denotes the AUROC performance when training with a subset of modalities. For instance, $v(\{t, c\})$ is the AUROC score of the model trained with clinical time series and CXR image ignoring the radiology notes. According to the idea of Shapley value, we calculate the marginal contribution of a certain modality by the difference in the model performance score when training with or without the modality. Therefore, $v(\{t, c, n\}) - v(\{t, c\})$ denotes the marginal contribution of radiology notes during the training process. The Shapley value of certain modality m is derived by a weighted sum of the marginal contribution of m in all possible subsets of modalities. It can be formulated as:

$$\phi_m(v) = \sum_{S \subseteq N - \{m\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!}(v(S \cup \{m\}) - v(S))$$

As the formula shows, the Shapley value can be seen as a weighted sum of all $v(S), S \subseteq N$. We calculate the Shapley value of modalities on the AUROC score respectively for different fusion strategies and tasks.

Shapley value has many properties, one of which is called the efficiency rule: $\sum_{m \in N} \phi_m(v) = v(N)$. This means that the Shapley values of all modalities add up to the total profit gained. Therefore, we can normalize the Shapley value of the modalities so that they add up to 1, and then the normalized Shapley value shows the proportion of contribution from each modality in the game and can be compared between trials with different fusion strategies and tasks.

Calculating the Shapley value needs the AUROC on all possible subsets of modality set N, including the empty set. We assume that the AUROC is 0 on the empty set. The normalized Shapley value of each modality on all three tasks is shown in Fig. 5.

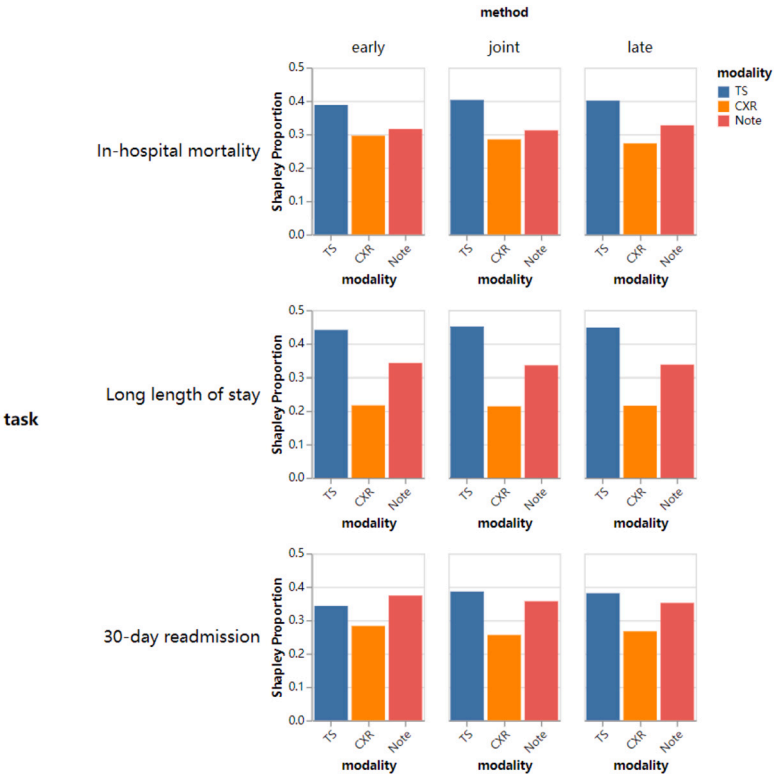

Fig. 5. Shapley value of each modality on the AUROC score in different task & fusion method trails. The subgraphs are arranged in a grid manner. The three rows correspond to three tasks while the three columns correspond to three fusion methods. The original Shapley value of the three modalities is normalized so that they add up to 1, which indicates the percentage contribution of the modality in the model performance that allows comparison between trials. Abbreviation in the figure: TS - Time Series, CXR - Chest X-ray.

4.6. Discussion

It is worth pointing out that F1 score, precision, recall, accuracy, AUROC, and AUPRC are reported for evaluation. The performance is highly task-related and relies on the distribution of the dataset. We can get several insights from the experiment results above.

Unimodal performance comparison. Among unimodal models, the model trained by clinical time series performs better than images and notes in all three tasks. Although get a slightly lower AUROC in 30-day readmission prediction, clinical time series performance surpasses the other two modalities by a considerable gap with other performance metrics. The possible reason is that clinical time series data contains vital features that can directly reflect the patient's health status. For example, an abnormal heart rate directly indicates poor health condition. On the other hand, CXR images alone may not be sufficient for accurately predicting mortality, long length of stay, and 30-day readmission. The reason might be that the CXR images only show the condition of the lungs and may not provide a comprehensive view of the patient's overall health status.

Multimodal performance boost. It is a general trend that models with multimodal inputs tend to earn higher AUROC and AUPRC scores than unimodal ones due to complementary information from multiple sources, even though the clinical time series partial model can get comparable F1 scores, precision, recall, and accuracy. Moreover, models with three modalities tend to earn higher performance than those with two modalities in many situations and have comparable results even if they are not the best. However, some modalities could be poisonous, such as notes in the early fusion setting of 30-day readmission prediction. The possible reason is that the notes are noisy which may confuse the model when further fine-tuning is disabled in the early fusion setting.

Fusion strategy comparison. Joint fusion outperforms early and late fusion in AUROC and AUPRC metrics, possibly due to its fine-tuned feature representation. Additionally, joint fusion has fewer parameters compared to late fusion, which aids in mitigating overfitting issues.

No single strategy dominates all performance metrics tested, indicating that there is no consistent trend in model performance. However, F1 score, precision, recall, and accuracy are sensitive to the classification threshold, while AUROC and AUPRC are threshold-agnostic. So we prefer joint fusion as it tends to be the best or at least comparable to the best on AUROC and AUPRC. Moreover, joint fusion has a more flexible architecture in that we can select different levels of features for fusion, endowing joint fusion potential to combine information from different modalities.

Modality contribution discussion. Fig. 5 shows that clinical time series variables contribute the most to the three tasks, and CXR contributes the least. The contribution distribution of modalities tends to be consistent across all three fusion methods for each

task, while slightly different for different tasks. For mortality prediction, clinical time series has a contribution of over 43% while CXR and Notes have less contribution of about 23% and 33%. For the long length of stay prediction, a contribution gap is also present and there is a 0.41-0.33-0.26 contribution distribution on TimeSeries-Note-CXR. For 30-day readmission, the contribution distribution varies a lot across different fusion strategies, but time series still surpasses the other two modalities. We believe that this is because a patient may just have one or two chest X-ray images and radiology notes, while tens of clinical variable observations during the admission. This frequency difference makes clinical time series more informative for patient's health status.

5. Conclusion

In this paper, we introduced a general framework that can integrate clinical time series, medical images, and clinical notes in EHR with 3 different fusion strategies and generate feature vectors for downstream predictive tasks. Performance on the three prediction tasks shows that extra modalities improve the performance on predictive tasks. Additionally, by calculating the contribution proportion of each modality with Shapley value, we found that clinical time series are the most helpful in the three tasks.

Note that the proposed framework can be readily adapted to fit both existing risk prediction models and tasks related to risk prediction. The framework is also compatible with more advanced fusion methods other than direct summation. For example, we can try weighted sum or tensor product to merge feature vectors from different modalities. It is also worth exploring to generate more fine-grained contribution explanations for variables and pixels from the input data samples.

We believe that the strengths of our model lie in the ability to combine different modalities and scalability to work with more additional modalities. However, one limitation of our current research is that we focus on the MIMIC-IV dataset, a single-site dataset, which may not be enough for representative patient distribution. Moreover, the MIMIC-IV dataset is large-scale and comprehensive, which provides abundant multimodal data to train our model. However, for the relatively small datasets, the performance of our framework may be affected by a higher rate of missing values and missing modalities.

Ethical statement

The MIMIC-IV, MIMIC-CXR, and MIMIC-IV-Note datasets are publicly available. The Institutional Review Board at the Beth Israel Deaconess Medical Center has reviewed the gathering of patient information for these datasets and the establishment of this research resource. They granted a waiver of informed consent and authorized the sharing of data.

CRediT authorship contribution statement

Yuanlong Wang: Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Data curation. **Changchang Yin:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation. **Ping Zhang:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The three datasets (MIMIC-IV, MIMIC-CXR, and MIMIC-IV-Note) used are all available at https://physionet.org/ with credentialed access. Our code for this project is available at https://github.com/Wang-Yuanlong/MultimodalPred.

Acknowledgements

This work was funded in part by the National Institutes of Health (NIH) under award number R01GM141279. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References

[1] L.L. Weed, Medical records that guide and teach (concluded), Yearb. Med. Inform. 212 (1968) 1.
[2] J. Henry, Y. Pylypchuk, T. Searcy, V. Patel, et al., Adoption of electronic health record systems among US non-federal acute care hospitals: 2008–2015, ONC data brief. 35, 2008–2015 2016.
[3] ONC, National Trends in Hospital and Physician Adoption of Electronic Health Records |, HealthIT. Gov., 2023.
[4] T. Sarwar, S. Seifollahi, J. Chan, X. Zhang, V. Aksakalli, I. Hudson, K. Verspoor, L. Cavedon, The secondary use of electronic health records for data mining: data characteristics and challenges, ACM Comput. Surv. 55 (2022) 1–40.
[5] A. Johnson, L. Bulgarelli, T. Pollard, S. Horng, L.A. Celi, R. Mark, MIMIC-IV (version 1.0), PhysioNet (2021), https://doi.org/10.13026/s6n6-xd98.
[6] A. Johnson, M. Lungren, Y. Peng, Z. Lu, R. Mark, S. Berkowitz, S. Horng, MIMIC-CXR-JPG - chest radiographs with structured labels (version 2.0.0), PhysioNet (2019), https://doi.org/10.13026/8360-t248.

[7] A.E. Johnson, T.J. Pollard, S. Berkowitz, N.R. Greenbaum, M.P. Lungren, C.Y. Deng, R.G. Mark, S. Horng, MIMIC-CXR: A large publicly available database of labeled chest radiographs, arXiv preprint arXiv:1901.07042, 2019 Jan 21.
[8] A. Johnson, T. Pollard, S. Horng, L.A. Celi, R. Mark, MIMIC-IV-Note: Deidentified free-text clinical notes (version 2.2), PhysioNet (2023), https://doi.org/10.13026/1n74-ne17.
[9] L.S. Shapley, Notes on the n-person game—ii: The value of an n-person game, 1951.
[10] E. Choi, M.T. Bahadori, J. Sun, J. Kulas, A. Schuetz, W. Stewart, RETAIN: an interpretable predictive model for healthcare using reverse time attention mechanism, Adv. Neural Inf. Process. Syst. 29 (2016), Curran Associates, Inc. Available at https://proceedings.neurips.cc/paper/2016/hash/231141b34c82aa95e48810a9d1b33a79-Abstract.html.
[11] E. Choi, M.T. Bahadori, E. Searles, C. Coffey, M. Thompson, J. Bost, J. Tejedor-Sojo, J. Sun, Multi-layer representation learning for medical concepts, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD'16, Association for Computing Machinery, New York, NY, USA, 2016, pp. 1495–1504.
[12] L. Rasmy, Y. Xiang, Z. Xie, C. Tao, D. Zhi, Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction, npj Digit. Med. 4 (2021) 1–13, https://doi.org/10.1038/s41746-021-00455-y.
[13] D. Zhang, C. Yin, J. Zeng, X. Yuan, P. Zhang, Combining structured and unstructured data for predictive models: a deep learning approach, BMC Med. Inform. Decis. Mak. 20 (2020) 280, https://doi.org/10.1186/s12911-020-01297-6.
[14] A. Ashfaq, A. Sant'Anna, M. Lingman, S. Nowaczyk, Readmission prediction using deep learning on electronic health records, J. Biomed. Inform. 97 (2019) 103256, https://doi.org/10.1016/j.jbi.2019.103256.
[15] M. Capan, P. Wu, M. Campbell, S. Mascioli, E.V. Jackson, Using electronic health records and nursing assessment to redesign clinical early recognition systems, Health Syst. 6 (2017) 112–121, https://doi.org/10.1057/hs.2015.19.
[16] F. Ma, R. Chitta, J. Zhou, Q. You, T. Sun, J. Gao, 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, Halifax NS Canada, ACM, 2017, pp. 1903–1911.
[17] Y. Li, S. Rao, J.R.A. Solares, A. Hassaine, R. Ramakrishnan, D. Canoy, Y. Zhu, K. Rahimi, G. Salimi-Khorshidi, BEHRT: transformer for electronic health records, Sci. Rep. 10 (2020) 7155, https://doi.org/10.1038/s41598-020-62922-y.
[18] J. Shang, T. Ma, C. Xiao, J. Sun, Pre-training of graph augmented transformers for medication recommendation, in: S. Kraus (Ed.), Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019, 2019, pp. 5953–5959, ijcai.org.
[19] A. Kline, H. Wang, Y. Li, S. Dennis, M. Hutch, Z. Xu, F. Wang, F. Cheng, Y. Luo, Multimodal machine learning in precision health: a scoping review, npj Digit. Med. 5 (2022) 171, https://doi.org/10.1038/s41746-022-00712-8.
[20] M. Golovanevsky, C. Eickhoff, R. Singh, Multimodal attention-based deep learning for Alzheimer's disease diagnosis, J. Am. Med. Inform. Assoc. 29 (2022) 2014–2022, https://doi.org/10.1093/jamia/ocac168.
[21] S.-C. Huang, A. Pareek, R. Zamanian, I. Banerjee, M.P. Lungren, Multimodal fusion with deep neural networks for leveraging CT imaging and electronic health record: a case-study in pulmonary embolism detection, Sci. Rep. 10 (2020) 22147, https://doi.org/10.1038/s41598-020-78888-w.
[22] Z. Yao, X. Hu, X. Liu, W. Xie, Y. Dong, H. Qiu, Z. Chen, Y. Shi, X. Xu, M. Huang, J. Zhuang, A machine learning-based pulmonary venous obstruction prediction model using clinical data and CT image, Int. J. Comput. Assisted Radiol. Surg. 16 (2021) 609–617, https://doi.org/10.1007/s11548-021-02335-y.
[23] R. Yan, F. Zhang, X. Rao, Z. Lv, J. Li, L. Zhang, S. Liang, Y. Li, F. Ren, C. Zheng, J. Liang, Richer fusion network for breast cancer classification based on multimodal data, BMC Med. Inform. Decis. Mak. 21 (2021) 134, https://doi.org/10.1186/s12911-020-01340-6.
[24] D. Nie, J. Lu, H. Zhang, E. Adeli, J. Wang, Z. Yu, L. Liu, Q. Wang, J. Wu, D. Shen, Multi-channel 3D deep feature learning for survival time prediction of brain tumor patients using multi-modal neuroimages, Sci. Rep. 9 (2019) 1103, https://doi.org/10.1038/s41598-018-37387-9, Number: 1 Publisher: Nature Publishing Group.
[25] L.R. Soenksen, Y. Ma, C. Zeng, L. Boussioux, K. Villalobos Carballo, L. Na, H.M. Wiberg, M.L. Li, I. Fuentes, D. Bertsimas, Integrated multimodal artificial intelligence framework for healthcare applications, npj Digit. Med. 5 (2022) 1–10, Publisher: Nature Publishing Group.
[26] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Adv. Neural Inf. Process. Syst. 30 (2017), Curran Associates, Inc. Available at https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html.
[27] S.M. Lundberg, G. Erion, H. Chen, A. DeGrave, J.M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, S.-I. Lee, From local explanations to global understanding with explainable AI for trees, Nat. Mach. Intell. 2 (2020) 56–67, https://doi.org/10.1038/s42256-019-0138-9.
[28] D. Fryer, I. Strümke, H. Nguyen, Shapley values for feature selection: the good, the bad, and the axioms, IEEE Access 9 (2021) 144352–144360, https://doi.org/10.1109/ACCESS.2021.3119110, Conference Name: IEEE Access.
[29] C. Yin, R. Liu, D. Zhang, P. Zhang, Identifying sepsis subphenotypes via time-aware multi-modal auto-encoder, in: R. Gupta, Y. Liu, J. Tang, B.A. Prakash (Eds.), KDD'20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020, ACM, 2020, pp. 862–872.
[30] Q.V. Le, T. Mikolov, Distributed representations of sentences and documents, in: Proceedings of the 31st International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014, in: JMLR Workshop and Conference Proceedings, vol. 32, 2014, pp. 1188–1196, JMLR.org. Available at http://proceedings.mlr.press/v32/le14.html.
[31] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (1997) 1735–1780, https://doi.org/10.1162/neco.1997.9.8.1735.
[32] F. Mohsen, H. Ali, N. El Hajj, Z. Shah, Artificial intelligence-based methods for fusion of electronic health records and imaging data, Sci. Rep. 12 (2022) 1–16, Publisher: Nature Publishing Group.
[33] A. Rajkomar, E. Oren, K. Chen, A.M. Dai, N. Hajaj, M. Hardt, P.J. Liu, X. Liu, J. Marcus, M. Sun, et al., Scalable and accurate deep learning with electronic health records, npj Digit. Med. 1 (2018) 18, Publisher: Nature Publishing Group UK London.
[34] J. Benbassat, M. Taragin, Hospital readmissions as a measure of quality of health care: advantages and limitations, Arch. Intern. Med. 160 (2000) 1074–1081, https://doi.org/10.1001/archinte.160.8.1074, https://jamanetwork.com/journals/jamainternalmedicine/articlepdf/415392/ira90007.pdf.