M3T-LM: A Multi-Modal Multi-Task Learning Model for Jointly Predicting Patient Length of Stay and Mortality

Junde Chen, Qing Li, Feng Liu, Yuxin Wen

***Abstract*—Ensuring accurate predictions of inpatient length of stay (LoS) and mortality rates is essential for enhancing hospital service efficiency, particularly in light of the constraints posed by limited healthcare resources. Integrative analysis of heterogeneous clinic record data from different sources can hold great promise for improving the prognosis and diagnosis level of LoS and mortality. Currently, most existing studies solely focus on single data modality or tend to single-task learning, i.e., training LoS and mortality tasks separately. This limits the utilization of available multi-modal data and prevents the sharing of feature representations that could capture correlations between different tasks, ultimately hindering the model's performance. To address the challenge, this study proposes a novel Multi-Modal Multi-Task learning model, termed as M3T-LM, to integrate clinic records to predict inpatients' LoS and mortality simultaneously. The M3T-LM framework incorporates multiple data modalities by constructing sub-models tailored to each modality. Specifically, a novel attention-embedded one-dimensional (1D) convolutional neural network (CNN) is designed to handle numerical data. For clinical notes, they are converted into sequence data, and then two long short-term memory (LSTM) networks are exploited to model on textual sequence data. A two-dimensional (2D) CNN architecture, noted as CRXMDL, is designed to extract high-level features from chest X-ray (CXR) images. Subsequently, multiple sub-models are integrated to formulate the M3T-LM to capture the correlations between patient LoS and modality prediction tasks. The efficiency of the proposed method is validated on the MIMIC-IV dataset. The proposed method attained a test MAE of 5.54 for LoS prediction and a test F_1 of 0.876 for mortality prediction. The experimental results demonstrate that our approach outperforms state-of-the-art (SOTA) methods in tackling mixed regression and classification tasks.**

***Index Terms*—Multi-task learning, Data-fusion model, Length of stay prediction, Deep learning.**

I. Introduction

HEALTHCARE systems continue to face a significant challenge of providing timely patient care while optimizing resource utilization, especially in the wake of the COVID-19 pandemic [1]. Inpatients' length of stay (LoS) and mortality are two crucial metrics that hospitals utilize to assess clinical quality and optimize resource allocation [2]. Prolonged LoS escalates the likelihood of encountering adverse events, such as poor nutritional levels, hospital-acquired infections, adverse drug events, and various other complications. Furthermore, prolonged LoS increases in the odds of inpatient mortality [3]. This has triggered hospitals to spend intensive efforts on resource allocation. Real-time demand capacity (RTDC) management [4] and multidisciplinary discharge rounds (MDRs) [5] have shown great promise as best practices in addressing these challenges, but their effectiveness relies on the accurate prediction of inpatients' LoS and mortality. With the rising prevalence of electronic health record (EHR) systems, patients' clinical records, such as patients' laboratory test results, vital signs, demographic information, clinical notes, and other details, are now accessible. Leveraging this abundant knowledge, sophisticated data-driven algorithms enable precise predictions for inpatients' LoS and mortality.

This research focuses on improving the service efficiency and management capabilities of hospitals by simultaneously predicting inpatient LoS and mortality. As mentioned previously, patients' LoS and mortality in a hospital are crucial indicators to assess the quality of care and effective allocation of healthcare resources. Therefore, predicting inpatient LoS (number of days) will be a regression prediction, and mortality will be a binary classification in this study. From the recent literature, it is evident that machine learning (ML) offers unprecedented opportunities to improve patient and clinical outcomes due to the great potential for learning essential features and extracting meaningful insights from data [6]. Some notable works include, but are not limited to, Gaussian Process Regression for clinical e-health modeling [7], prediction of Intensive Care Unit (ICU) LoS based on four ML methods such as Logistic Regression (LR), Support Vector Machine (SVM), Random forest (RF), and XGBoost [8], a Hierarchical Attention Network (HAN) for LoS and mortality predictions [9], ensemble learning for improving predictive performance [10], U-Net-Based Models for medical image segmentation [11], CNN for medical image classification [12], etc. However, the majority of existing ML models in healthcare either rely exclusively on a single data modality or solely for a single task [13]. With the increasing availability and accessibility of multi-modal data, multi-modal deep learning (DL) models [14], aiming to integrate data of different distributions, sources, and formats into a unified space where both inter-modality and cross-modality aspects can be uniformly captured, have been successful in a wide range of domains, such as autonomous driving and video classification through combining visual features from cameras along with data from Light Detec-

Corresponding author: Yuxin Wen

J. Chen and Y. Wen are with the Dale E. and Sarah Ann Fowler School of Engineering, Chapman University, Orange, CA 92866 USA (e-mail:jundchen@chapman.edu; yuwen@chapman.edu).

Q. Li is with Department of Industrial and Manufacturing Systems Engineering, Iowa State University, Ames, IA 50011 USA (e-mail:qlijane@iastate.edu).

F. Liu is with School of Systems and Enterprises, Stevens Institute of Technology, Hoboken, NJ 07030 USA (e-mail:fliu22@stevens.edu).

tion and Ranging (LiDAR) sensors [15], emotion recognition through the fusion of audiovisual content with textual users' comments [16], and process monitoring in manufacturing using multimodal sensor data [17]. The main challenge of multi-modal data fusion is that data from different sources and file formats exhibit heterogeneity and high-dimensionality, seldom adhering to uniformity, and this is especially the case with clinical data. The complex nature of clinical data imposes significant challenges on how to efficiently make joint representations of heterogeneous modalities in a way that enables their seamless integration. Consequently, even with significant importance, the predictions of inpatient LoS and mortality using multi-modal data have received less attention in the literature [2].

Another noticeable trend is that most clinical machine learning systems focus on single clinical prediction tasks. Nonetheless, in the real-world clinical environment, multiple tasks always demonstrate interdependence. For instance, while the risk of heart disease and the likelihood of diabetes development represent distinct medical conditions, they share underlying physiological factors such as blood pressure, cholesterol levels, and family medical history [18]. Multi-task learning (MTL), a subfield of machine learning, fosters the interchange of insights among interconnected tasks by training multiple related tasks simultaneously using a single model. By sharing information between related tasks, MTL improves the generalization and performance of the model by leveraging the shared information of related tasks. Le et al. [19] proposed a convolutional neural network (CNN) based multi-task classification and segmentation architecture for cancer diagnosis in mammography. Yu et al. [20] used a multi-task recurrent neural network with an attention mechanism to predict patient mortality in hospitals. Despite the achievements in medical predictions using MTL, there has not been much effort to simultaneously incorporate multi-modal clinical data and multi-task learning with the aim of enhancing prediction performance. Tan et al. [21] proposed a multi-modal and multi-task DL framework called MultiCoFusion to combine the power of different modalities and tasks for cancer prognosis prediction. Their experimental results indicate that the joint learning of multiple tasks can utilize the intrinsic association between features (i.e., genes), and thus, can further promote the learning performance. However, they manually extracted features from histopathological images and mRNA expression data, and LoS prediction was not their research topic. Harerimana et al. [9] developed a hierarchical deep attention model to forecast the LoS and in-hospital mortality from ICD codes and demographic data. Unfortunately, the LoS was predicted in a classification manner. In addition, LoS and mortality tasks were trained separately.

To address the aforementioned challenges, in this study, we propose a novel Multi-Modal Multi-Task learning model, termed as M3T-LM, to perform the LoS regression and mortality classification tasks simultaneously. Multiple data modalities, including demographic data, clinical notes, laboratory test results, and medical images, are integrated to be used in our scheme. According to different data modalities, the basic models (sub-models) are constructed using relevant data types. Concretely, a novel attention-embedded one-dimensional (1D) CNN is designed to handle numerical data. By converting the texts to sequence data, two long short-term memory (LSTM) networks are used to model on clinical notes. A two-dimensional (2D) CNN architecture, named CRXMDL, is designed to extract high-level features from chest X-ray (CXR) images. Subsequently, multiple sub-models are integrated to form the M3T-LM to capture the correlations between patient LoS and modality prediction tasks. It is important to note that predicting inpatients' LoS and mortality involves a challenging mixed-task scenario, encompassing both regression and classification tasks. A novel predictive framework is proposed to address this challenge. Overall, the key contributions of this study can be recapitulated as follows:

- A joint classification-regression scheme that implements mixed-task types using heterogeneous data modalities is proposed to predict inpatients' LoS and mortality simultaneously.
- An enhanced squeeze-and-excitation (SE)-block, where the 2D pooling layer is replaced by a 1D one and two non-linear fully-connected layers are substituted by a 1×1 convolution layer to address numerical data and decrease the number of parameters, is incorporated into the network for adaptive feature calibration.
- CXR images are an integral component of our scheme, where we have developed an innovative CNN model referred to as CRXMDL. This model utilizes the InceptionResNet V2 as its backbone network, known for its powerful feature extraction capabilities. We further enhance this architecture by embedding three convolution blocks with 32, 16, and 8 filters of size 3×3, a max pooling (MAP) layer, a flatten layer, and a fully connected (FC) layer. These additions are designed to capture and leverage the most salient features from CXR images, thereby improving the accuracy and robustness of our predictions.
- A unified model that incorporates losses from both regression and classification tasks is developed. An adaptive loss weight assignment solution is proposed to determine the optimal weights for these tasks automatically, enhancing the model's overall performance.

The rest of this paper is structured as follows. Section II briefly introduces the relevant work and identifies the research gaps. Section III discusses the proposed methodology in detail. Section IV presents experiment results as well as comparative analysis. Finally, Section V concludes the paper and points out future work.

II. Related work

First, we conduct a review of the literature that focuses on the related studies on inpatients' LoS and mortality prediction. Then, a review of the multi-modal multi-task learning in the healthcare field is presented. Subsequently, the current research gaps are discussed in this section.

A. Length of stay prediction

Accurate prediction of LoS can increase patient satisfaction by reducing unnecessary wait times and saving hospital costs. The existing ML models for LoS prediction can be broadly grouped into two categories: classification models and regression models [22]. In classification models, the aim is to group the LoS into multiple classes, e.g., short stay, medium stay, and long stay, based on the number of days that the patient stays in the hospital. Morton et al. [23] categorized the LoS of diabetic inpatients into long-term and short-term stays. Multiple classification models such as SVM, RF, and LASSO-based multi-task learning were used in their comparison experiments. Although the SVM and RF achieved the desired results, they believe that multi-task learning is promising for LoS prediction. Thompson et al. [24] divided the LoS of newborns into prolonged LoS or not, and they used well-known ML algorithms such as LR, Decision Tree (DT), SVM, RF, and neural networks (NN) to implement the class prediction. The RF, DT, and NN achieved impressive performance. Nevertheless, recent studies have pointed out that the LoS distributions exhibit a significant right-skewed pattern [9], [25], which indicates that the balance of the dataset is disrupted, with only a limited number of instances demonstrating long LoS. Consequently, this imbalance causes the model to treat classes with long LoS as anomalies, leading to a decrease in classification accuracy. Therefore, it is more appropriate to formulate the LoS task as a regression problem. Modeling LoS prediction as a regression problem has gained less attention in the literature. Tsai et al. [26] applied a NN method to predict LoS for cardiology patients. They concluded that the NN model is robust for predicting prolonged LoS. However, there is still room for improvement in the accuracy of their model. Using a cluster-boosted regression method, Rouzbahman et al. [27] conducted mortality and LoS predictions for ICU inpatients. Their findings indicated enhanced accuracy in regression predictions for both mortality and LoS. However, determining an optimal number of clusters remains challenging and involves a degree of subjectivity. Muhlestein et al. [10] trained an ML ensemble model to predict inpatient LoS after brain tumor surgery. Their experimental results demonstrated the good performance of a ML ensemble model for LoS prediction. However, the ML ensemble model integrates multiple sub-models and enhances the complexity of calculation.

B. Inpatient mortality prediction

Accurate prediction of inpatient mortality plays a vital role in evaluating disease severity, interventions, assessing the efficacy of novel therapies, and guiding healthcare initiatives. Over the past few decades, great efforts have been invested in the prediction of inpatient mortality. Ruzicka et al. [28] applied XGBoost, to predict patients' mortality in hospitals, and compared it with a traditional unregularized LR model. In their experiments, the XGBoost outperformed the LR but was not competitive with existing methods. Ganapathy et al. [29] compared several models, such as LR, Binary Discriminant Analysis (BDA), Bayesian Linear Regression (BLR), NN, and RF, for inpatient mortality prediction, and the BLR model achieved the best precision in their experiments. They concluded that the ML classifiers had the best predictive ability in comparison to statistical models. Caicedo-Torres et al. [30] designed a deep learning model called ISeeU2 to predict mortality inside the ICU. Their proposed model outperformed the compared baselines, highlighting the valuable insights that can be extracted from raw nursing notes. Similarly, in another research, Zeng et al. [31] proposed a recurrent neural network (RNN)-based DL architecture to predict the mortality for all admissions in the ICU. Although their proposed approach outperforms classical machine learning methods such as LR, RF, and XGBoost, the issue of imbalanced positive and negative sample distributions remains unaddressed. Using image-Transformed electrocardiograms (ECG) waveforms, Kondo et al. [32] conducted short-term mortality prediction for cardiac care unit patients, and their method successfully reached the desired prediction accuracy. Nevertheless, it is noteworthy that the model solely relies on image data, which limits the overall accuracy of their approach.

C. Multi-Modal Multi-Task learning

Compared to single modality models, multi-modal models have the capacity of producing more reliable results owing to their ability to perceive different aspects of the data, leading to enhanced model accuracy and reliability. Multi-task learning (MTL), which can solve multiple learning tasks at the same time, while exploiting commonalities and differences across tasks, has proven to be more reliable in identifying related characteristics, less sensitive to data noises, and less overfitting risk [33]. Zhang et al. [34] proposed a 3-dimensional multi-task DL model based on MLP-Mixer architecture to simultaneously implement FDG/AV45-PET SUVR and AD status prediction tasks in Alzheimer's Disease Diagnosis. Their experimental results show that MTL can share feature representations, which is beneficial for both tasks. Shao et al. [35] proposed a multi-task multi-modal learning method for joint prognosis and diagnosis of cancer patients. Two types of data including histopathological images and genomic data were used in their scheme to address both prognosis and diagnosis tasks. They concluded that the MTL captured the correlation between different tasks and obtained better performance than single-task learning. Using different data modalities such as histopathological images and mRNA expression data, Tan et al. [21] built a multi-modal multi-task learning model to perform the survival analysis and grade classification for cancer prognosis diagnosis. They concluded that using multi-modal data would perform better than using only single-modal data. In another study, Liu et al. [36] proposed jointly identifying brain diseases and predicting clinical scores using both magnetic resonance imaging (MRI) and patient demographic information. Their experimental findings demonstrate that the MTL outperforms the single-task learning.

Although some joint learning models have been proposed, certain models incorporate only a single data modality. Moreover, most of them first extract hand-crafted features from images and pre-process the data separately, and the separate process might lack effective coordination, consequently resulting in suboptimal learning performance. Besides, most research implemented the same task type but not mixed-task types. To address abovementioned research gap, in this paper, we establish a multi-modal multi-task learning model to simultaneously implement mixed-type regression and classification tasks using multiple data modalities. Specifically, we aim to simultaneously predict inpatients' length of stay and mortality as they have proven to be closely related for the inpatients after ICU admission, and share common feature representations needed to train regression and classification models. Heterogeneous medical data modalities, including but not limited to, static numerical data (demographics, healthcare examination), unstructured texts (clinical notes, long procedure texts), and Chest X-ray images, are used by the proposed multi-modal multi-task model (M3T-LM) to implement the automatic prediction of inpatients' LoS and mortality.

III. Methodology

The proposed M3T-LM method includes the following key distinctions: (1) To maximize the utilization of available data, a multi-modal data fusion that fuses diverse data modalities, including patient demographic information, diagnosis, free clinical notes, laboratory test results, and medical images, is implemented in our scheme. (2) A multi-task learning model with a shared network layer is proposed to capture the correlations between inpatients' LoS and modality prediction tasks, since these two tasks are intrinsically associated with each other. (3) Mixed-task types are learned in our scheme. Different from the same task type implemented in most existing research, the mixed-task types including the regression and classification tasks are simultaneously performed for the inpatient LoS and mortality prediction. The proposed approach leverages the interconnections among the diverse data and tasks, which potentially improves model efficiency and reduces overfitting risk through modeling nonlinear within and cross-modality relationships. Fig. 1 provides the flow diagram of the proposed procedure. In the following, we first present the architecture of the M3T-LM, and then the optimization process is discussed in detail.

A. Architecture of the M3T-LM

Define the input data as $\boldsymbol{D}$ and is composed of diverse subsets $\boldsymbol{D} = \{\boldsymbol{D_1}, \boldsymbol{D_2}, ..., \boldsymbol{D_M}\}$, where M is the total number of modalities (As shown in Table II, three different data modalities including numerical, text, and image data are used in our scheme). Denote the length of stay for each subject as $y_i,\ i = 1, 2, , \dots,\ N$, labels of C categories as $z^c = \{z_n^c\}_{n=1}^N$, N is the number of total samples. First, the sub-models are built based on the different data modalities to learn abstractions of the data from raw data directly. Specifically, for the numerical data, a novel attention-embedded 1D CNN noted Att-1DCNN is developed to extract meaningful information. Using 32 convolutional kernels with the size of 3, two cascaded 1D convolution layers followed by a 1D max pooling layer are used to extract favorable features. Then, a modified squeeze-and-excitation (SE)-block, where the 2D pooling layer is replaced by a 1D one and two non-linear fully-connected (FC) layers are substituted by a 1×1 convolution layer to address 1D numerical data and decrease the number of parameters, is incorporated into the network for adaptive feature calibration [37]. Next, following the SE-block, a spatial attention (SA) mechanism that can help CNN extract global features via mining the inter-spatial relationship between features is embedded into the network to infer the importance of spatial points [38]. By this means, the features obtained by the attention mechanism and bottom convolution block are fused to generate the output of the Att-1DCNN, which comprehensively extracts the useful information of numerical data for the prediction tasks of the model.

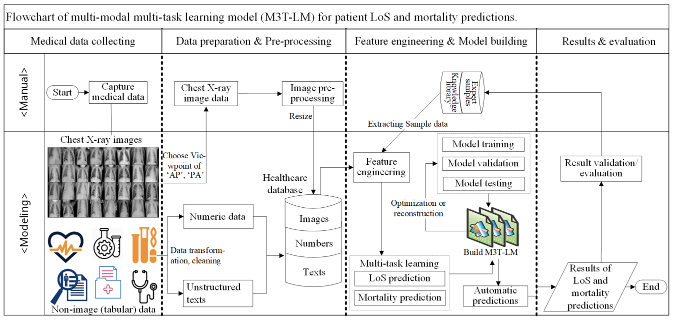


Fig. 1. The flowchart of the proposed procedure.

For unstructured texts, the long short-term memory (LSTM) network is an effective and end-to-end DL method for text processing. Besides, word embedding is a popular technique to map words or phrases from vocabulary to a corresponding vector of continuous values. However, directly modeling sequential notes using word embeddings and DL can be time-consuming. It may not be practical since the length of different documents varies. Therefore, the tokenizer is first employed to implement word segmentation for long texts. Then, a Text2Seq function [39] is used to transform the text data into the sequence variables. To capture the dependence along with sequence variables, two LSTM networks are separately designed to take the output of Text2Seq for d_icd_diagnoses and d_icd_procedures long titles to infer the sequence-dependent feature representations. Here, the hyper-parameter of the perceptron number is set to 2 and L2 regularization is employed for suppressing overfitting risk.

The integration of clinical data and chest X-ray images showed a favorable impact on the predictive performance in prognostication tasks, and it also delivered a positive performance for in-hospital mortality prediction and phenotype classification [12], [14], [40]. Particularly, the effects of CXR images were illustrated by Hayat et al. [40], who observed a significant improvement in accuracy when using the CXR images along with clinical data to build a multi-modal fusion model. Therefore, we further integrate the CXR images into our modeling scheme. For the CXR image data, we devise a convolutional neural network architecture named CXRMDL, in which the InceptionResnet V2 [41] is used as a backbone network of the model. Transfer learning is applied in our modeling scheme. Specifically, the InceptionResnet V2 is

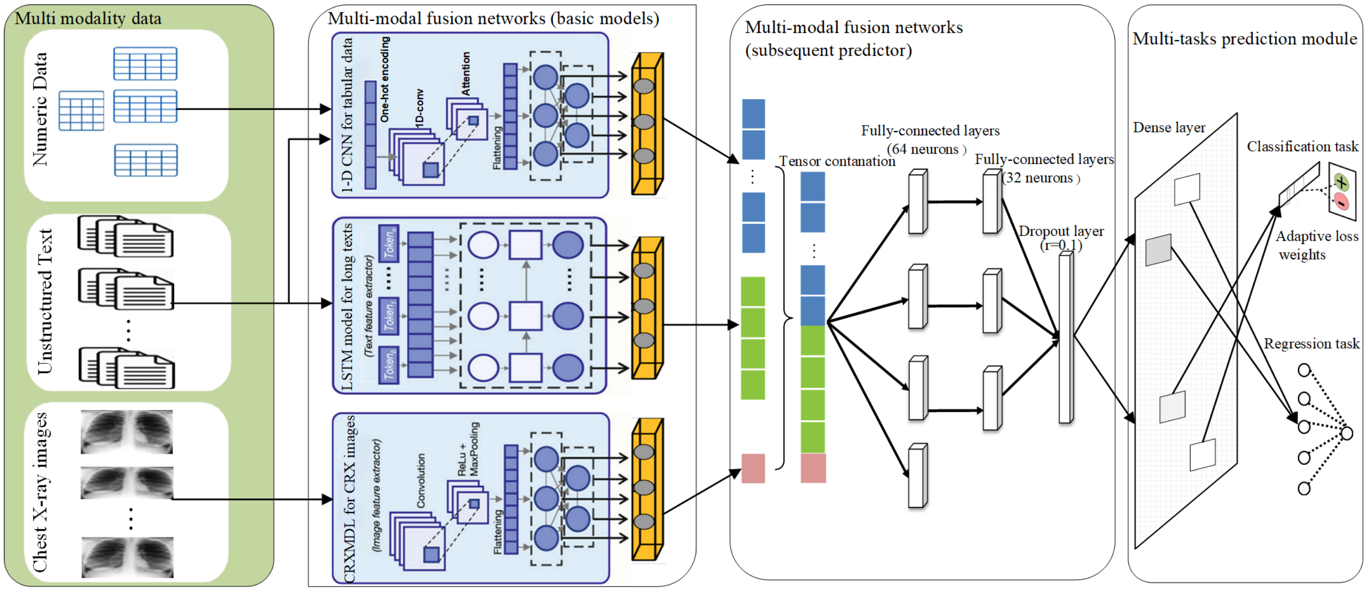


Fig. 2. The overall architecture of the proposed M3T-LM.

TABLE I
THE MAJOR PARAMETERS OF THE PROPOSED MODEL.

Layer (module)	Input Shape	Filter No.	Kernel Size	Output Shape	Params	Repeat
num (InputLayer)	(None, 26, 1)	-	-	(None, 26, 1)	-	1
conv1d (Conv1D)	(None,26,1)	32	3	(None, 24, 32)	128	1
conv1d-1 (Conv1D)	(None,24,32)			(None, 22, 32)	3,104	1
MaxPooling1D	(None,22,32)	-	-	(None,11,32)	-	1
GlobalAveragePooling1D	(None,11,32)	-	-	(None,32)	-	1
conv1d-2 (Conv1D)	(None,32,1)	32	3	(None,32,1)	3	1
conv1d-3 (Conv1D)	(None,1,11,2) (None,1,1,32)	32	3	(None,1,11,1)	3	1
sigmoid-1	(None,1,11,1) (None,1,11,32)	-	-	(None,1,11,1)	-	1
ad (InputLayer)	(None,23)	-	-	(None,23)	-	1
ap (InputLayer)	(None,18)	-	-	(None,18)	-	1
multiply_1 (Multiply)	(None,1,11,32) (None,1,11,1)	-	-	(None,1,11,32)	-	1
eth (InputLayer)	(None,1)	-	-	(None,1)	-	1
Flatten	(None,1,11,32) (None,27,1)	-	-	(None,352) (None,27)	-	2
embedding	(None,1) (None,18) (None,23)	-	-	(None,1,32)	192+16,960+17,120	3
cxg (InputLayer)	(None,128,128,1)	-	-	[(None,128,128,1)]	-	1
LSTM	(None,23,32) (None,18,32)			(None,2)	280	2
ohe (InputLayer)	(None,28)	-	-	(None,28)	-	1
flatten (Flatten)	(None,1,32)	-	-	(None, 32)	-	1
sequential (Sequential)	(None,128,128,1)	-	-	(None, 1536)	54,336,160	1
Concatenate layer	(None,378) (None,28) (None, 32) (None, 1536) (None, 2)	-	-	(None, 1978)	-	2
Dense (drop=0.1)	(None,1978)	-	-	(None, 32)	126,656+2,080	2
Regression / Class	(None,32)	-	-	(None,1) (None,2)	33+33	1
Total	-	-	-	-	54,503,032 (33 layers)	-

employed with the truncated top layers, which is followed by a global average pooling (GAP) layer, a flatten layer, and a batch normalization (BN) layer to extract deep-level features of CXR images.

Subsequently, the multiple basic sub-models including Att-1DCNN, LSTM, and CXRMDL are integrated to form a novel Multi-Modal Multi-Task learning model, where two densely-connected (DC) layers with the neuron numbers of 64 and 32 are embedded into the networks to change the vector dimensions. Following each DC layer, a dropout layer with a dropout rate of 0.1 is added to suppress the overfitting risks. After that, a fully-connected (FC) ReLu layer and a FC Softmax layer are used in the network for the final LoS prediction and mortality classification tasks. In brief, the crucial steps are summarized below.

(1) The raw dataset is pre-processed and augmented to $\boldsymbol{D}$, which is divided into diverse subsets $\boldsymbol{D} = \{\boldsymbol{D_1}, \boldsymbol{D_2}, ..., \boldsymbol{D_M}\}$. Where $\boldsymbol{D_k} = \{(x_{k,1}, y_{k,1}, z_{k,1})), (x_{k,2}, y_{k,2}, z_{k,2}), ..., (x_{k,M}, y_{k,M}, z_{k,M})\}$, $x_{k,m}$ indicates the extracted feature data, and $(y_{k,m}, z_{k,m})$ denotes the corresponding target variables (e.g., LoS and mortality), $k, m \in \{1, 2, ..., M\}$.

(2) Construct the backbone network models $H = \{H_1, H_2, ..., H_M\}$, which is used for addressing data in different modalities such as numerical data, text data, and CXR image data. Each backbone model is fed corresponding data subsets $\boldsymbol{D} = \{\boldsymbol{D_1}, \boldsymbol{D_2}, ..., \boldsymbol{D_M}\}$ extracted in the pre-processed stage, where the data transformation and data cleaning are implemented.

(3) The outputs of basic models are concatenated and used as the input to the subsequent (secondary) predictor F. Here, the subsequent predictor F consists of 2 DC layers with 64 and 32 neurons, 2 dropout layers, a FC ReLu layer, and then a FC Softmax layer is used for the final LoS regression and mortality classification tasks, respectively. Fig. 2 portrays an overall architecture of the proposed M3T-LM, and the major parameters are summarized in Table I.

B. Optimization of the M3T-LM

In the proposed framework, each branch of the joint model learns a different task, and therefore it is necessary to specify a loss function for each task. In this research, the inpatient LoS and mortality predictions are modeled as the regression and classification tasks, respectively. Thus, the regression and classification loss functions in the proposed M3T-LM are separately defined below.

(1) Regression loss function. In general, the mean squared error (MSE) function is the most-used loss function employed in deep learning models for addressing regression problems, while it also has some demerits, such as sensitivity to outliers. Therefore, to reduce the impact of singularity values, we introduce a custom regression loss function [42] in the network for the LoS prediction task. The formula of the regression loss function is defined by:

$$L_{reg} = \begin{cases} \frac{1}{N}\sum\limits_{i=1}^{N} \frac{1}{2}(y_i - \hat{y}_i), for\, |y_i - \hat{y}_i| \leq \delta, \\ \frac{1}{N}\sum\limits_{i=1}^{N} (\delta\, |y_i - \hat{y}_i| - \frac{1}{2}\delta^2), otherwise. \end{cases} \tag{1}$$

where y_i and $\hat{y}_i$ denote the actual and predicted values, respectively. δ is a hyperparameter of the threshold value, and here it is set to 2 according to the hyperparameter tuning results.

(2) Classification loss function. The in-hospital mortality prediction belongs to a classification problem, and the Binary Cross Entropy (BCE) loss function is exploited in our network to address the in-hospital mortality prediction task. The formula of BCE loss function is expressed as:

$$L_{class} = -\frac{1}{N}\sum_{i=1}^{N} y_i \cdot log(p(y_i)) + (1 - y_i) \cdot log(1 - p(y_i)) \tag{2}$$

where y_i represents the actual class and $p(y_i)$ denotes the predicted probability of that class. As a consequence, the two different loss functions are established for multi-modal multi-task learning. However, only the result of one loss function can be updated in the process of backpropagation, so a joint loss function must be defined to integrate the two different loss functions, and the weighted sum method is the most commonly used scheme. The weighted total loss function is formulated as:

$$L_{total}^{(t)} = \sum\nolimits_{i=1} w_i^{(t)} L_i^{(t)} \tag{3}$$

where w_i denotes the weight for the i-th loss function L_i, and t implies the t-th epoch of training. The performance of the system is highly relied on the defined weights between each task's loss, but tuning these weights by hand is a great challenge and an expensive process. Therefore, we include the loss weights in the definition of the loss function itself and develop an adaptive way to update loss weights through callbacks, which manage the changes internally. The loss weight update can be defined as:

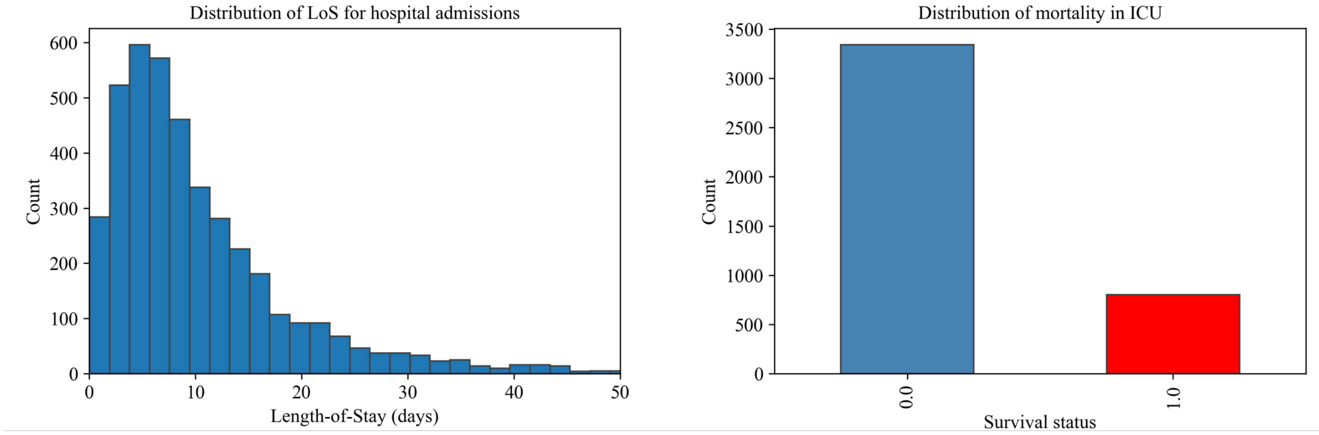
$$w_i^{(t+1)} \leftarrow w_i^{(t)} - \lambda \nabla_{w_i} L_{grad} \tag{4}$$

where λ is a constant hyper-parameter, and L_{grad} denotes the gradient loss, which is introduced especially to depict the loss caused by loss weight w. The formula of gradient loss L_{grad} is written by:

$$L_{grad}(t, w_i^{(t)}) = \sum\nolimits_{i} \left| G_i^{(t)} - \overline{G^{(t)}} \times \left[r_i^{(t)} \right]^{\alpha} \right|_1 \tag{5}$$

$$G_i^{(t)} = \left\| \nabla_\theta w_i^{(t)} L_i^{(t)} \right\|_2, r_i^{(t)} = \frac{L_i^{(t)} / L_i^{(0)}}{E_{task}\left[L_i^{(t)} / L_i^{(0)} \right]} \tag{6}$$

In Eq. (5), $G_i^{(t)}$ is the value of gradient normalization on the i-th task in the t-th epoch of training, which is calculated by the L2 norm of the weighted loss gradient. $\overline{G^{(t)}}$ represents the mean of gradient normalization for all tasks in the t-th epoch of training. $r_i^{(t)}$ denotes the relative training speed of the i-th task in the t-th epoch of training, which is calculated as the ratio of the training speed of the i-th task to the average training speed of all tasks. Overall, the loss weight is regarded as an optimization parameter in this solution, and the L_{grad} of loss weight w is established in each epoch of update. The initial weights of the regression and classification loss functions are both set to 0.5, and the gradient update is



(a) The distribution of inpatient LoS

(b) Distribution of mortality

Fig. 3. The distribution of LoS and mortality.

implemented for each epoch of training.

IV. Experiments

In this section, we present the empirical performance evaluation of the proposed approach for LoS and mortality prediction tasks. We first evaluate the accuracy of the proposed M3T-LM compared to state-of-the-art (SOTA) methods. Next, we perform the hyperparameter optimization and fine-tuning via the random search for optimal sets of essential hyperparameters to maximize the prediction performance of the model. Ultimately, we assess the efficacy of fused data modalities and newly added modules for the proposed approach via ablation study.

A. Dataset Description and Preprocessing

MIMIC, short for the Medical Information Mart for Intensive Care, is a large database of clinical records for patients admitted to the Beth Israel Deaconess Medical Center (BIDMC). The MIMIC-IV, which consists of comprehensive clinical information on hospital stays for patients, contains de-identified records of 50,048 individual patients admitted to the ICU or emergency department (ED) at the BIDMC in Boston, MA, USA, between 2008 and 2019. The MIMIC-IV's most recent version (v1.0) [43], which was released on June 22, 2022, improves on MIMIC-III [44] to provide public access to the EHR data based on the BIDMC's MetaVision clinical information system. Whilst, MIMIC-CXR-JPG v2.0.0 [45], which is a large image dataset comprised of 227,827 CXR images sourced from the BIDMC between 2011 and 2016.

TABLE II
The characteristics of the data.

Characteristic	Data type	No. of variables	Name of variables
Demographic variables	Numerical & Categorical	6	Subject_id, gender, anchor_age, anchor_year, anchor_year_group, dod
Chart Event variables	Numerical & Categorical	9	Hadm_id, stay_id, charttime, storetime, itemid, value, valuenum valueuom, warning
Laboratory Event variables	Numerical & Categorical	23	Labevent_id, hadm_id, specimen_id, itemid, charttime, storetime, value, valuenum, valueuom, ref_range_lower,ref_range_upper, flag, priority; WBC count, neutrophils count, monocytes count, lymphocytes count, platelets count, hemoglobin, glucose, chloride, creatinine, BUN
Procedure Event variables	Numerical & Categorical	10	Patient weight, total amount, total amount uom, isopenbag, continue in next dept, cancel reason, status description, comments_date, original amount, original rate
Text Note variables	Text	2	AdmitDiagnosis, AdmitProcedure
Chest X-ray variables	Image	1	Chest X-ray image

This dataset is freely available to facilitate and encourage broad research in the field of medical computer vision.

To efficiently predict the inpatients' LoS and mortality, we utilize the MIMIC-IV v1.0 dataset combined with the MIMIC-CXR-JPG v2.0.0 dataset in this study. All the data are de-identified, where patient identifiers are removed according to the Health Insurance Portability and Accountability Act (HIPAA) Safe Harbor provision. The MIMIC-IV v1.0 database includes a wide range of patient records, such as patients' demographic information, laboratory test results, procedures and diagnoses, free-text notes authored by clinicians, medication orders, etc. The tables in the MIMIC-IV v1.0 dataset mainly contain ADMISSIONS, DIAGNOSIS_ICD, D_DIAGNOSIS_ICD PATIENTS, ICDSTAYS, PROCEDURES_ICD, and D_PROCEDURES_ICD. Among them, the ADMISSIONS table provides records for each hospitalization including each patient's admission and discharge time and the source of the admission. The DIAGNOSIS_ICD table gives the diagnosis category information. The PATIENTS table provides timing information and demographics for each patient, and the ICDSTAYS table provides the ICU data for each hospital admission. The PROCEDURES_ICD table presents the procedure code for inpatients and the corresponding procedure names are included in the D_PROCEDURES_ICD table. Additionally, for the patients' chest X-ray images, they are stored in the MIMIC-CXR-JPG database in JPG format with structured labels. The characteristics of the data we used are summarized in Table II. In this research, the LoS is defined as the time between hospital discharge and admission measured in days. The mortality is depicted by the field of hospital_expired_flag in the table ADMISSIONS, where 1 indicates the death and 0 indicates survival of patients in hospitals. Data preprocessing, including data cleaning, data transformation, revision of outliers, interpolation of missing data [46], [47] are implemented for the original tabular data. As a result, a total of 51 variables, including blood, circulatory, digestive, endocrine, injury, and nervous, are extracted from the MIMIC-IV v1.0 dataset. For the CXR image data, only the images with the ViewPosition of "PA (Posterior-Anterior)" or "AP (Anterior-Posterior)" are chosen in our experiments since they are photographed from the front view. As such, 4,144 CXR image samples are used for the LoS and mortality prediction experiments. Fig. 3 portrays the distribution of LoS and mortality. From Fig. 3 it can be visualized that most LoS is under 20 days, and there is a class-imbalanced problem in the distribution of mortality. The survival category (class 0) comprises the majority of samples, while the mortality category (class 1) consists of only a small number of samples. The distribution is extremely unbalanced. To cope with this challenge, the Synthetic Minority Oversampling Technique (SMOTE) [48] is utilized to augment the minority class samples to ensure a balanced distribution of positive and negative samples in the training set. Using the SMOTE, new synthetic data are generated to make the number of samples in the mortality category very close to that in the survival category.

B. Experiment setup and performance metrics

The experiments are conducted using Python 3.6 deep learning framework, where the commonly-used libraries including Keras, Scikit-learn, TensorFlow, and Matplotlib are utilized with the aid of a graphics processing unit (GPU). The hardware environment for operating the Python DL framework to implement the proposed M3T-LM contains the AMD EPYC 7502P 32-Core Processor, 32 GB memory, and RTX A6000 GPU.

To evaluate the performance of LoS prediction, the standard measure metrics like the mean absolute error (MAE), root mean square error ($RMSE$), coefficient of determination (R-$Square$ or R^2), and explained variance (E_{VAR}) [6], [8], [14] are utilized, which are calculated by

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{7}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{8}$$

$$R^2 = 1 - \sum_{i=1}^{N}(y_i - \hat{y}_i)^2 / \sum_{i=1}^{N}(y_i - \bar{y})^2 \tag{9}$$

$$E_{VAR} = 1 - var(\boldsymbol{y} - \hat{\boldsymbol{y}})/var(\boldsymbol{y}) \tag{10}$$

where $\hat{y}_i$, y_i, and $\bar{y}$ indicate the predicted value, actual value, and mean of actual values, respectively. N is the number of total samples, and $var(\cdot)$ implies the variance function. MAE and $RMSE$ reflect the mean of the absolute error and the square root of the average squared error between the predicted value and actual value, respectively. R^2 measures the proportion of the dependent variable change that can be interpreted by the independent variable, and the E_{VAR} reveals the explanatory power of models. For both the E_{VAR} and R^2, the ideal value is equal to 1, while greater values are worse for the MAE and $RMSE$ indicators. Moreover, widely-used metrics including $Accuracy$ (Acc), $Precision$ (Pre), $Recall$ (Rec), and $F1$-$Score$ ($F1$) [32] are utilized to investigate the efficiency of mortality prediction, which can be calculated by the following equations:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{11}$$

$$Precision = \frac{TP}{(TP + FP)} \tag{12}$$

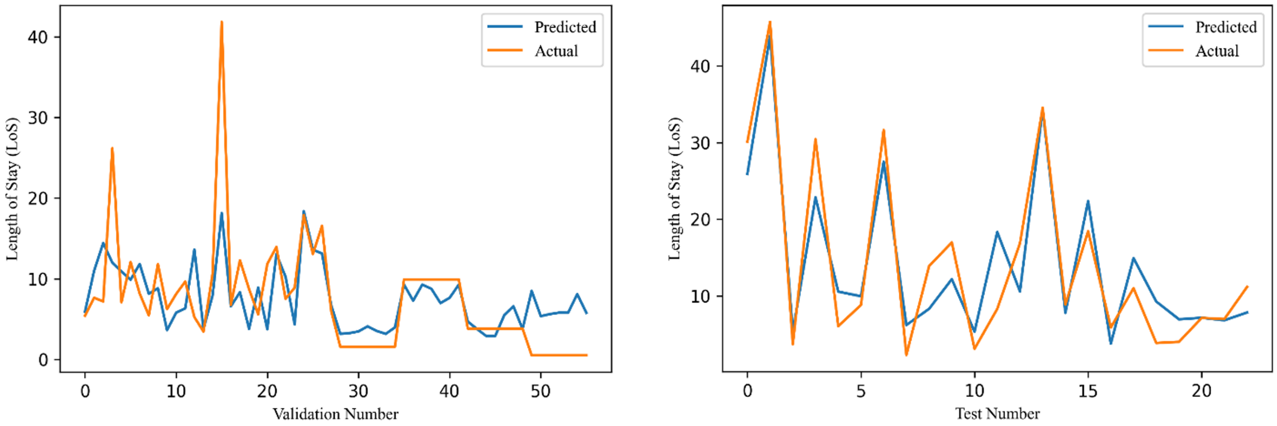
$$Recall = \frac{TP}{(TP + FN)} \tag{13}$$

$$F1 = \frac{(2 \times Precision \times Recall)}{(Precision + Recall)} \tag{14}$$

where TP is true positive, FP is false positive, TN is true negative, and FN is false negative.

C. Results and Discussion

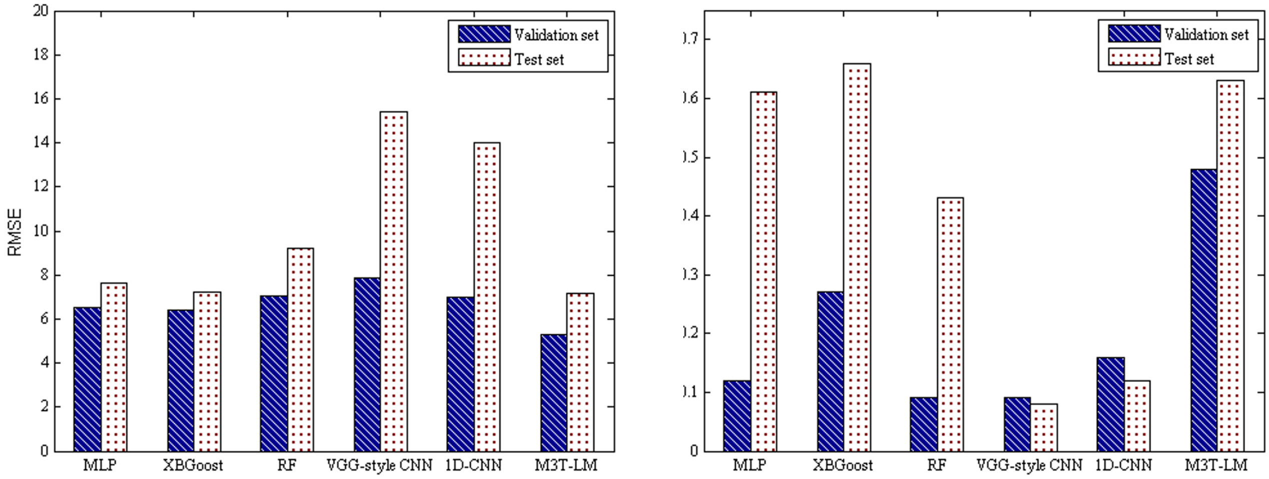
To demonstrate the robustness of the proposed method, the mostly-used ML methods, multilayer perceptron (MLP),



(a) Prediction comparison on the validation set

(b) Prediction comparison on the test set

Fig. 4. The LoS prediction effect of the proposed approach.



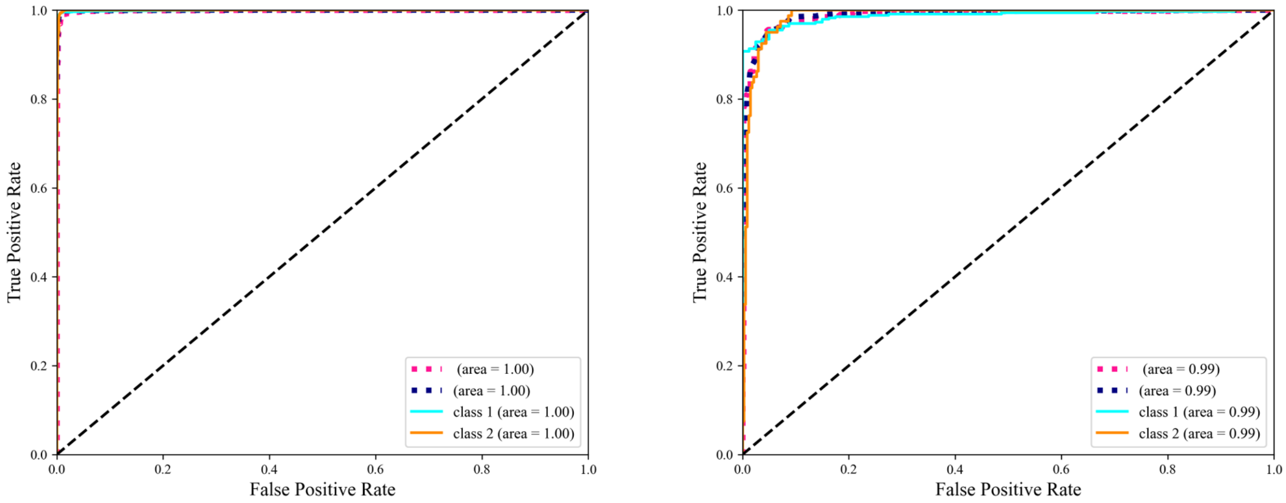
(a) *RMSE*

(b) *R-Square*

Fig. 5. The $RMSE$ and E_{VAR} comparison of different methods.

TABLE III
LoS PREDICTION OF DIFFERENT METHODS.

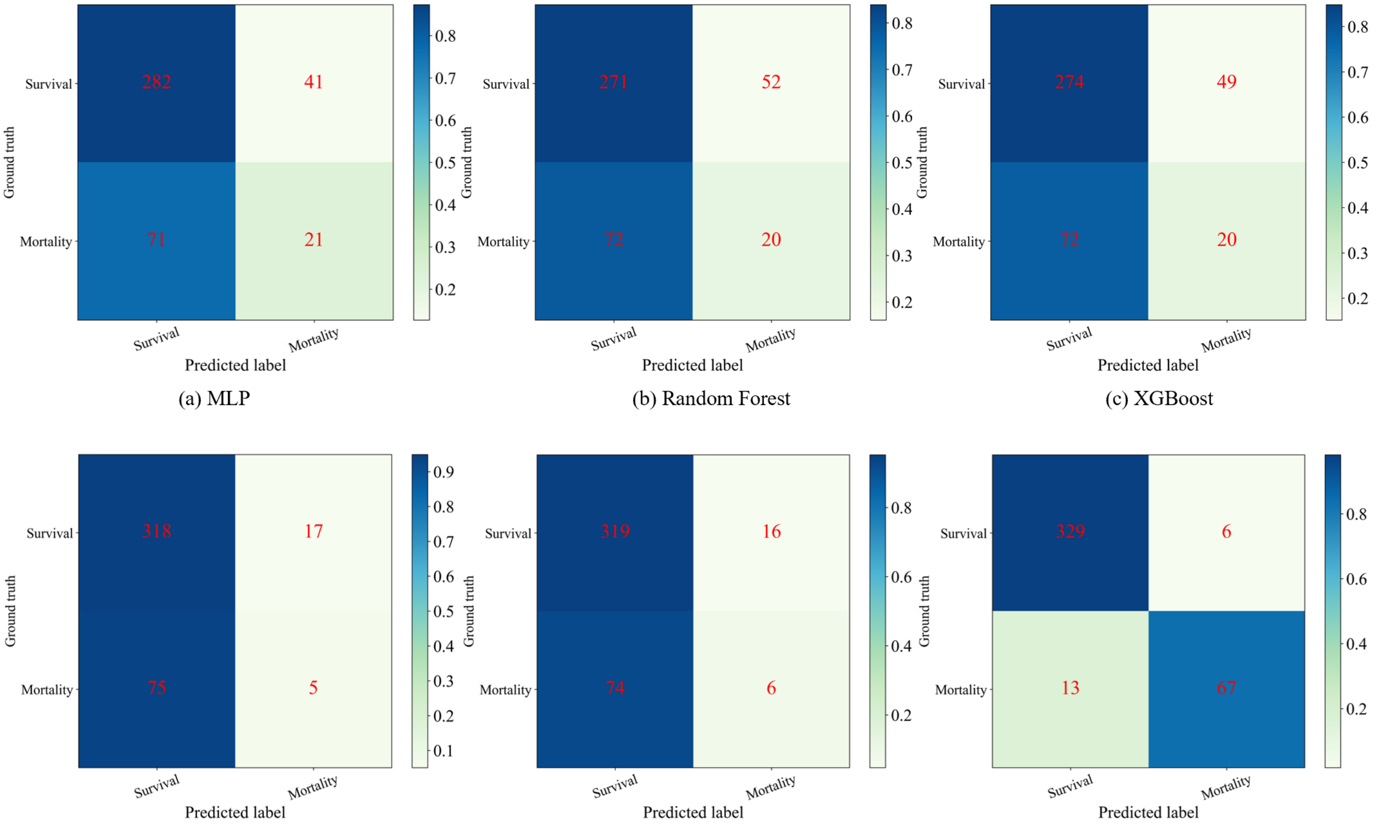
No.	Methods	Training set				Validation set				Test set				Time
		MAE	$RMSE$	R^2	E_{VAR}	MAE	$RMSE$	R^2	E_{VAR}	MAE	$RMSE$	R^2	E_{VAR}	
1	MLP	5.33	10.28	0.45	0.45	4.63	6.54	0.09	0.12	5.65	7.64	0.58	0.61	0:00:40
2	XGBoost	3.43	5.23	0.85	0.85	5.04	6.40	0.13	0.27	5.01	7.21	0.62	0.66	0:00:31
3	RF	4.55	8.05	0.66	0.66	0.66	7.04	-0.05	0.09	6.28	9.22	0.38	0.43	0:00:31
4	VGG-style CNN	7.52	15.16	-0.18	0.07	5.10	7.87	-0.30	0.09	10.62	15.41	-0.71	0.08	0:02:46
5	1D-CNN	6.76	14.18	-0.03	0.14	4.28	6.97	-0.02	0.16	9.40	14.02	-0.41	0.12	0:03:04
6	M3T-LM	3.68	10.44	0.44	0.44	3.85	5.30	0.41	0.48	5.54	7.18	0.62	0.63	0:34:22



(a) ROC curves on the training dataset

(b) ROC curves on the test dataset

Fig. 6. The ROC curve of mortality prediction.



(a) MLP

(b) Random Forest

(c) XGBoost

(d) VGG-style CNN

(e) 1D-CNN

(f) Proposed M3T-LM

Fig. 7. The test confusion matrices of different methods.

TABLE IV
MORTALITY PREDICTION OF DIFFERENT METHODS.

No.	Methods	Training set (%)				Validation set (%)				Test set (%)				Time (s)
		Acc	Pre	Rec	$F1$	Acc	Pre	Rec	$F1$	Acc	Pre	Rec	$F1$	
1	MLP	91.28	85.74	98.42	91.64	91.44	85.25	99.46	91.81	73.01	33.87	22.82	27.27	0:00:36
2	XGBoost	88.60	83.98	94.56	88.96	88.96	83.72	95.74	89.33	70.84	28.98	21.73	24.84	0:00:37
3	RF	89.11	82.10	99.21	89.85	90.76	84.13	99.64	91.23	70.12	27.77	21.73	24.39	0:00:36
4	VGG-style CNN	98.32	98.31	98.26	98.29	95.59	95.51	95.51	95.51	77.83	22.72	6.25	9.80	0:01:58
5	1D-CNN	98.24	98.77	97.62	98.19	94.91	96.76	92.75	94.71	78.31	27.27	7.50	11.76	0:03:08
6	M3T-LM	98.91	97.84	100.00	98.90	98.72	97.80	99.65	98.71	95.42	91.78	83.75	87.58	0:27:23

TABLE V
COMPARISON WITH STATE-OF-THE-ART METHODS.

ID	References	Year	Description	$RMSE$ (LoS)	R^2 (LoS)	$F1$ (mortality)
1	Vaswani et al. [49]	2017	Transformer	6.18	0.27	0.738
2	Harutyunyan et al. [50]	2019	LSTM	6.61	0.28	0.745
3	Ma et al. [51]	2020	ConCare	-	-	0.778
4	Rocheteau et al. [52]	2021	Temporal Pointwise Convolution (TPC)	5.20	0.59	0.784
5	Al-Dailami et al. [2]	2022	Temporal Dilated Separable Convolution with Context Aware Feature Fusion (TDSC-CAFF)	4.30	0.64	0.821
6	Shu et al. [53]	2023	ML-based scoring models	-	-	0.613
7	This study	2024	M3T-LM	7.18	0.62	0.876

RF, and extreme gradient boosting (XGBoost), along with a VGG-style CNN and one-dimensional CNN (1D-CNN) are selected for comparative analysis. Different from the proposed approach that fits multi-modal data, the classical ML methods can only take a single data modality, such as numerical data. Therefore, these compared methods are conducted on the tabular data of the MIMIC-IV v1.0 dataset. To ensure a fair comparison, the core hyperparameters of the compared models are set to the same as that of the proposed approach. Specifically, the mini-batch size is set to 64, with a learning rate of 1×10^{-3}, 30 training epochs and the RMSprop [54] optimizer. The dataset is randomly divided into training, validation, and test sets in a 7:2:1 ratio. We employ the leave-one-out cross-validation approach for performance evaluation, where 90% of the samples are used for training and validation, and the remaining 10% for testing. Fig. 4 illustrates the LoS prediction performance of the proposed approach on randomly selected samples from both the validation and test datasets. In Fig. 4, the orange curve represents the actual values of inpatients' LoS and the blue curve denotes the predicted LoS. It can be seen from Fig. 4 that the predicted values are very close to their actual values for most samples, indicating the efficacy of the proposed approach.

Table III presents the overall LoS prediction performance of different methods. Fig. 5 visualizes the $RMSE$ and E_{VAR} comparison of different methods. It can be seen from Table III that the proposed approach realizes the R^2 of 0.62 and 0.41, and the $RMSE$ of 7.18 and 5.30 on the test set and validation set, respectively, which are superior to that of other comparison methods. The proposed approach achieves the best results. Notably, although the ensemble learning algorithms such as XGBoost and RF, perform better than the proposed M3T-LM in the training set, a significant decline in validation and test performance is observed for both XGBoost and RF, which has also been shown in Fig. 5. It is noted that the proposed M3T-LM takes 34 minutes for 30 epochs of training, which is also reported in Table III. Due to the large number of parameters in the proposed deep learning framework and the concurrent execution of two tasks, the proposed model requires more time than benchmark methods. Though the time consumption of the proposed approach is slightly higher than that of other compared methods, this aspect remains manageable and can be further improved by various optimization techniques.

Next, we evaluate the performance of the proposed approach for the inpatients' mortality prediction. Fig. 6 depicts the receiver operating characteristic (ROC) curves of the proposed approach, and the test confusion matrices of different methods are portrayed in Fig. 7. Table IV presents the overall mortality prediction performance of different methods. As depicted in Fig. 6, the proposed approach exhibits superior operating characteristics, with the ROC curves of all categories positioned close to the top-left corner of the figure. This positioning signifies the validity and effectiveness of the proposed approach for mortality

TABLE VI

LoS PREDICTION RESULTS WITH HYPERPARAMETER OPTIMIZATION.

Mini-batch size	lr=0.001				lr=0.002				lr=0.005				Time (s)
	MAE	$RMSE$	R^2	E_{VAR}	MAE	$RMSE$	R^2	E_{VAR}	MAE	$RMSE$	R^2	E_{VAR}	
32	4.23	6.31	0.38	0.40	4.61	6.47	0.35	0.35	4.81	6.97	0.25	0.25	0:52:02
64	5.54	7.18	0.62	0.63	4.55	6.30	0.38	0.39	4.59	6.88	0.26	0.30	0:27:23
128	4.62	6.87	0.27	0.37	4.72	6.40	0.36	0.38	5.19	7.37	0.15	0.16	0:14:21
256	4.41	6.38	0.37	0.38	4.64	7.05	0.23	0.38	10.22	12.99	-1.62	0.02	0:08:28

TABLE VII

THE RESULTS OF ABLATION EXPERIMENTS.

Ablation approach	Test accuracy of LoS prediction				Accuracy of mortality prediction (%)				Time (s)
	MAE	$RMSE$	R^2	E_{VAR}	Acc	Pre	Rec	$F1$	
Delete images (CXRMDL)	6.33	10.14	0.39	0.44	89.39	100.00	45.00	62.06	0:04:56
Delete long texts (LSTM)	6.94	10.56	0.36	0.39	60.96	33.05	100.00	49.69	0:24:12
Delete attention module	5.84	9.37	0.37	0.42	92.53	72.47	98.75	83.59	0:26:52
This study	5.54	7.18	0.62	0.63	95.42	91.78	83.75	87.58	0:27:23

prediction. In addition, it can be observed from the confusion matrix of Fig. 7(f) that the M3T-LM has accurately identified most of the samples. The 67 mortality samples have been correctly recognized by the proposed approach except for 13 misidentified samples. Likewise, in addition to 6 samples misclassified into the mortality category, the 329 survival samples have all been correctly identified by the proposed approach. As a consequence, the proposed approach achieves a test $Accuracy$ of 95.42%, and the test $Precision$, $Recall$, and $F1$-$Score$ have also realized no less than 91.78%, 83.75%, and 87.58% respectively, as presented in Table IV.

Moreover, we conduct a performance evaluation of the proposed method in comparison to the findings presented in the latest literature concerning the prediction of LoS and mortality as shown in Table V. From Table V it can be visualized that the proposed approach delivers a comparable result and outperforms most of the existing methods on the MIMIC-IV v1.0 datasets. In summary, the outcomes of the comparative analysis affirm the excellence of the proposed method in predicting both LoS and mortality.

D. Hyperparameter optimization

In this section, we implement a grid search for optimal sets of essential hyperparameters including mini-batch size (bs) and learning rate (lr) on the prediction of inpatient LoS. The range of the mini-batch size hyperparameter is set as $(|B|) \in \{32, 64, 128, 256\}$, and the learning rate $(lr) \in \{0.001, 0.002, 0.005\}$. We train our model using hyperparameters from these sets for 30 epochs on the publicly available MIMIC-IV v1.0 dataset with the same splits, as mentioned in Section IV C. We found the best hyperparameter set for the LoS prediction is a mini-batch size of 64 with $lr = 0.001$. Table VI presents the prediction performance of the proposed method with different hyperparameter settings.

E. Ablation study

To gain a deeper understanding of the sub-models and different modalities contributing to a system's performance, ablation study on our model is performed.

Table VII summarizes the ablation experiment results. In the first ablation experiment, we remove the usage of the CXR image data modality and delete the CXRMDL module in our model. We notice a major decrease in the test accuracy, where the MAE and $RMSE$ in LoS prediction rise to 6.33 and 10.14 (increase by 0.79 and 2.96), and the R^2 and E_{VAR} drop to 0.39 and 0.44 (decrease by 0.23 and 0.19). The $Accuracy$, $Recall$, and $F1$-$Score$ in mortality prediction drop to 89.39%, 45.00%, and 62.06% (decrease by 6.03%, 38.75%, and 25.52%), respectively. On another front, it is noted that the time consumption of this ablation model shows a significant decrease from 27 minutes 23 seconds to 4 minutes 56 seconds (a reduction of over 22 minutes). This ablation experiment results demonstrate that removing the CXR image data modality has a great impact on the performance compared to the multi-modal data aggregation model. Subsequently, we remove the long text data modality addressed by the LSTM model in our networks. We notice that a significant drop in accuracy occurs in this ablation model. The test MAE and $RMSE$ in LoS prediction rise to 6.94 and 10.56 (increase by 1.40 and 3.38), and the test R^2 and E_{VAR} drop to 0.36 and 0.39 (decrease by 0.26 and 0.24), respectively. Likewise, the test $Accuracy$, $Precision$, and $F1$-$Score$ in mortality prediction separately drop to 60.96%, 33.05%, and 49.69% (decrease by 34.46%, 58.73%, and 37.89%). Consequently, though the efficacy of the ablated models is still better than that of other compared baselines, it suffers a notable decline in comparison with the multi-modal data aggregation model proposed in the study. In the second ablation experiment, we remove the newly added attention module from the networks to investigate the performance of the proposed method. We notice a minor drop in accuracy occurs in this ablation model, where the test MAE and $RMSE$ of the ablated model separately rise to 5.84 and 9.37 (increase by 0.30 and 2.19) in LoS prediction. The test $Accuracy$, $Precision$, and $F1$-$Score$ of the proposed method in mortality prediction also drop to 92.53%, 72.47%, and 83.59% (decrease by 2.89%, 19.31%, and 3.99%), respectively. This ablation experiment demonstrates that the results of the model adding the attention mechanism are slightly better than that of the model without the attention module, and removing the attention module has a minor negative influence on the model accuracy.

V. Conclusion

Estimating the inpatient LoS and mortality accurately is a challenging daily task in the field of health care. This study proposes a novel Multi-Modal Multi-Task learning model called M3T-LM to predict patient outcomes, specifically, remaining LoS and inpatient mortality. Leveraging mixed regression and classification tasks, M3T-LM simultaneously predicts inpatient LoS and mortality from multi-modal data. Acknowledging the skewed distribution of LoS, the proposed M3T-LM treats LoS prediction as a regression task, delivering more informative results by estimating the actual number of days rather than assigning classes. At the same time, M3T-LM integrates mortality prediction, recognizing its close association with inpatient LoS scenarios. The two tasks share standard feature representations necessary for the mixed-task model training. The main advantage of the proposed method is its capability of utilizing the inherent correlation within multiple task types to guide the feature selection process, which can further promote the learning performance. Besides, multiple data modalities are effectively utilized by the proposed method in a unified model, which leads to more effective resource allocation, higher prognostic accuracy, and better informative clinical decision-making. Impressively, experimental findings demonstrate that the proposed M3T-LM is superior to other SOTA baseline methods on both tasks.

While the proposed approach yields satisfactory results, it has some limitations related to computational complexity. In the future, we plan to incorporate model pruning algorithms to simplify the model and enhance its efficiency. Another interesting direction is that, in response to the escalating concerns regarding data privacy and security, we plan to incorporate privacy-preserving techniques into our model to ensure the safeguarding of sensitive information while maintaining effective data fusion without harming the model predictive performance.

Acknowledgment

This research is partially funded by the National Science Foundation under Grant No. 2246158.

References

[1] J. Sheng, J. Amankwah-Amoah, Z. Khan, and X. Wang, "Covid-19 pandemic in the new era of big data analytics: Methodological innovations and future research directions," *British Journal of Management*, vol. 32, no. 4, pp. 1164–1183, 2021.

[2] A. Al-Dailami, H. Kuang, and J. Wang, "Predicting length of stay in icu and mortality with temporal dilated separable convolution and context-aware feature fusion," *Computers in Biology and Medicine*, vol. 151, p. 106 278, 2022.

[3] A. H. Association *et al.*, "Aha hospital statistics: Fast facts on us hospitals," *American Hospital Association, available at: www/aha/org (accessed May 31, 2017)*, 2017.

[4] R. Resar, K. Nolan, D. Kaczynski, and K. Jensen, "Using real-time demand capacity management to improve hospitalwide patient flow," *The Joint Commission Journal on Quality and Patient Safety*, vol. 37, no. 5, 217–AP3, 2011.

[5] N. Meo, E. Paul, C. Wilson, J. Powers, M. Magbual, and K. M. Miles, "Introducing an electronic tracking tool into daily multidisciplinary discharge rounds on a medicine service: A quality improvement project to reduce length of stay," *BMJ open quality*, vol. 7, no. 3, e000174, 2018.

[6] I. T. Peres, S. Hamacher, F. L. C. Oliveira, F. A. Bozza, and J. I. F. Salluh, "Data-driven methodology to predict the icu length of stay: A multicentre study of 99,492 admissions in 109 brazilian units," *Anaesthesia Critical Care & Pain Medicine*, vol. 41, no. 6, p. 101 142, 2022.

[7] L. Clifton, D. A. Clifton, M. A. Pimentel, P. J. Watkinson, and L. Tarassenko, "Gaussian processes for personalized e-health monitoring with wearable sensors," *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 1, pp. 193–197, 2012.

[8] L. Hempel, S. Sadeghi, and T. Kirsten, "Prediction of intensive care unit length of stay in the mimic-iv dataset," *Applied Sciences*, vol. 13, no. 12, p. 6930, 2023.

[9] G. Harerimana, J. W. Kim, and B. 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," *Journal of Biomedical Informatics*, vol. 118, p. 103 778, 2021.

[10] W. E. Muhlestein, D. S. Akagi, J. M. Davies, and L. B. Chambless, "Predicting inpatient length of stay after brain tumor surgery: Developing machine learning ensembles to improve predictive performance," *Neurosurgery*, vol. 85, no. 3, p. 384, 2019.

[11] R. Yousef, S. Khan, G. Gupta, *et al.*, "U-net-based models towards optimal mr brain image segmentation," *Diagnostics*, vol. 13, no. 9, p. 1624, 2023.

[12] A. W. Salehi, S. Khan, G. Gupta, *et al.*, "A study of cnn and transfer learning in medical imaging: Advantages, challenges, future scope," *Sustainability*, vol. 15, no. 7, p. 5930, 2023.

[13] S.-C. Huang, A. Pareek, S. Seyyedi, I. Banerjee, and M. P. Lungren, "Fusion of medical imaging and electronic health records using deep learning: A systematic review and implementation guidelines," *NPJ digital medicine*, vol. 3, no. 1, p. 136, 2020.

[14] J. Chen, Y. Wen, M. Pokojovy, *et al.*, "Multi-modal learning for inpatient length of stay prediction," *Com-*

puters in Biology and Medicine, vol. 171, p. 108 121, 2024.

[15] D. Feng, C. Haase-Schütz, L. Rosenbaum, *et al.*, "Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 3, pp. 1341–1360, 2020.

[16] S. Nemati, R. Rohani, M. E. Basiri, M. Abdar, N. Y. Yen, and V. Makarenkov, "A hybrid latent space data fusion method for multimodal emotion recognition," *IEEE Access*, vol. 7, pp. 172 948–172 964, 2019.

[17] J. Petrich, Z. Snow, D. Corbin, and E. W. Reutzel, "Multi-modal sensor fusion with machine learning for data-driven process monitoring for additive manufacturing," *Additive Manufacturing*, vol. 48, p. 102 364, 2021.

[18] L. Men, N. Ilk, X. Tang, and Y. Liu, "Multi-disease prediction using lstm recurrent neural networks," *Expert Systems with Applications*, vol. 177, p. 114 905, 2021.

[19] T.-L.-T. Le, N. Thome, S. Bernard, V. Bismuth, and F. Patoureaux, "Multitask classification and segmentation for cancer diagnosis in mammography," *arXiv preprint arXiv:1909.05397*, 2019.

[20] R. Yu, Y. Zheng, R. Zhang, Y. Jiang, and C. C. Poon, "Using a multi-task recurrent neural network with attention mechanisms to predict hospital mortality of patients," *IEEE journal of biomedical and health informatics*, vol. 24, no. 2, pp. 486–492, 2019.

[21] K. Tan, W. Huang, X. Liu, J. Hu, and S. Dong, "A multi-modal fusion framework based on multi-task correlation learning for cancer prognosis prediction," *Artificial Intelligence in Medicine*, vol. 126, p. 102 260, 2022.

[22] Z. Lu, W. Chang, S. Meng, *et al.*, "The effect of high-flow nasal oxygen therapy on postoperative pulmonary complications and hospital length of stay in postoperative patients: A systematic review and meta-analysis," *Journal of Intensive Care Medicine*, vol. 35, no. 10, pp. 1129–1140, 2020.

[23] A. Morton, E. Marzban, G. Giannoulis, A. Patel, R. Aparasu, and I. A. Kakadiaris, "A comparison of supervised machine learning techniques for predicting short-term in-hospital length of stay among diabetic patients," in *2014 13th International Conference on Machine Learning and Applications*, IEEE, 2014, pp. 428–431.

[24] B. Thompson, K. O. Elish, and R. Steele, "Machine learning-based prediction of prolonged length of stay in newborns," in *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, IEEE, 2018, pp. 1454–1459.

[25] J. Chen, T. Di Qi, J. Vu, and Y. Wen, "A deep learning approach for inpatient length of stay and mortality prediction," *Journal of Biomedical Informatics*, vol. 147, p. 104 526, 2023.

[26] P.-F. J. Tsai, P.-C. Chen, Y.-Y. Chen, *et al.*, "Length of hospital stay prediction at the admission stage for cardiology patients using artificial neural network," *Journal of healthcare engineering*, vol. 2016, 2016.

[27] M. Rouzbahman, A. Jovicic, and M. Chignell, "Can cluster-boosted regression improve prediction of death and length of stay in the icu?" *IEEE journal of biomedical and health informatics*, vol. 21, no. 3, pp. 851–858, 2016.

[28] D. Ruzicka, T. Kondo, G. Fujimoto, A. P. Craig, S.-W. Kim, and H. Mikamo, "Development of a clinical prediction model for recurrence and mortality outcomes after clostridioides difficile infection using a machine learning approach," *Anaerobe*, vol. 77, p. 102 628, 2022.

[29] S. Ganapathy, K. Harichandrakumar, P. Penumadu, K. Tamilarasu, and N. S. Nair, "Comparison of bayesian, frequentist and machine learning models for predicting the two-year mortality of patients diagnosed with squamous cell carcinoma of the oral cavity," *Clinical Epidemiology and Global Health*, vol. 17, p. 101 145, 2022.

[30] W. Caicedo-Torres and J. Gutierrez, "Iseeu2: Visually interpretable mortality prediction inside the icu using deep learning and free-text medical notes," *Expert Systems with Applications*, vol. 202, p. 117 190, 2022.

[31] G. Zeng, J. Zhuang, H. Huang, *et al.*, "Use of deep learning for continuous prediction of mortality for all admissions in intensive care units," *Tsinghua Science and Technology*, vol. 28, no. 4, pp. 639–648, 2023.

[32] T. Kondo, A. Teramoto, E. Watanabe, *et al.*, "Prediction of short-term mortality of cardiac care unit patients using image-transformed ecg waveforms," *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 11, pp. 191–198, 2023.

[33] D. D. Solomon, S. Khan, S. Garg, *et al.*, "Hybrid majority voting: Prediction and classification model for obesity," *Diagnostics*, vol. 13, no. 15, p. 2610, 2023.

[34] Z.-C. Zhang, X. Zhao, G. Dong, and X.-M. Zhao, "Improving alzheimer's disease diagnosis with multi-modal pet embedding features by a 3d multi-task mlp-mixer neural network," *IEEE Journal of Biomedical and Health Informatics*, 2023.

[35] W. Shao, T. Wang, L. Sun, *et al.*, "Multi-task multi-modal learning for joint diagnosis and prognosis of human cancers," *Medical image analysis*, vol. 65, p. 101 795, 2020.

[36] M. Liu, J. Zhang, E. Adeli, and D. Shen, "Joint classification and regression via deep multi-task multi-channel learning for alzheimer's disease diagnosis," *IEEE Transactions on Biomedical Engineering*, vol. 66, no. 5, pp. 1195–1206, 2018.

[37] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[38] C. Chen, T. Wang, Y. Liu, L. Cheng, and J. Qin, "Spatial attention-based convolutional transformer for bearing remaining useful life prediction," *Measurement Science and Technology*, vol. 33, no. 11, p. 114 001, 2022.

[39] J. Brownlee, "How to prepare text data for deep learning with keras," *Machine Learning Mastery, Disponível em: https://machinelearningmastery. com/prepare-text-data-*

deep-learning-keras/.[acesso em: 15 de nov. de 2020], 2019.

[40] N. Hayat, K. J. Geras, and F. E. Shamout, "Medfuse: Multi-modal fusion with clinical time-series data and chest x-ray images," in *Machine Learning for Healthcare Conference*, PMLR, 2022, pp. 479–503.

[41] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 31, 2017.

[42] Q. Wang, Y. Ma, K. Zhao, and Y. Tian, "A comprehensive survey of loss functions in machine learning," *Annals of Data Science*, pp. 1–26, 2020.

[43] A. E. Johnson, L. Bulgarelli, L. Shen, *et al.*, "Mimic-iv, a freely accessible electronic health record dataset," *Scientific data*, vol. 10, no. 1, p. 1, 2023.

[44] A. E. Johnson, T. J. Pollard, L. Shen, *et al.*, "Mimic-iii, a freely accessible critical care database," *Scientific data*, vol. 3, no. 1, pp. 1–9, 2016.

[45] A. E. Johnson, T. J. Pollard, N. R. Greenbaum, *et al.*, "Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. arxiv 2019," *arXiv preprint arXiv:1901.07042*, 2019.

[46] C. Fan, M. Chen, X. Wang, J. Wang, and B. Huang, "A review on data preprocessing techniques toward efficient and reliable knowledge discovery from building operational data," *Frontiers in energy research*, vol. 9, p. 652 801, 2021.

[47] A. Kiourtis, A. Mavrogiorgou, G. Manias, and D. Kyriazis, "Ontology-driven data cleaning towards lossless data compression," in *Challenges of Trustable AI and Added-Value on Health*, IOS Press, 2022, pp. 421–422.

[48] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[49] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[50] H. Harutyunyan, H. Khachatrian, D. C. Kale, G. Ver Steeg, and A. Galstyan, "Multitask learning and benchmarking with clinical time series data," *Scientific data*, vol. 6, no. 1, p. 96, 2019.

[51] L. Ma, C. Zhang, Y. Wang, *et al.*, "Concare: Personalized clinical feature embedding via capturing the healthcare context," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, 2020, pp. 833–840.

[52] E. Rocheteau, P. Liò, and S. 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.

[53] T. Shu, J. Huang, J. Deng, *et al.*, "Development and assessment of scoring model for icu stay and mortality prediction after emergency admissions in ischemic heart disease: A retrospective study of mimic-iv databases," *Internal and Emergency Medicine*, vol. 18, no. 2, pp. 487–497, 2023.

[54] T. Tieleman and G. Hinton, "Rmsprop: Divide the gradient by a running average of its recent magnitude. coursera: Neural networks for machine learning," *COURSERA Neural Networks Mach. Learn*, vol. 17, 2012.