DuKA: A Dual-Keyless-Attention Model for Multi-modality EHR Data Fusion and Organ Failure Prediction

Zhangdaihong Liu, Xuan Wu, Yang Yang, David A. Clifton

***Abstract*— Objective: Organ failure is a leading cause of mortality in hospitals, particularly in intensive care units. Predicting organ failure is crucial for clinical and social reasons. This study proposes a dual-keyless-attention (DuKA) model that enables interpretable predictions of organ failure using electronic health record (EHR) data. Methods: Three modalities of medical data from EHR, namely diagnosis, procedure, and medications, are selected to predict three types of vital organ failures: heart failure, respiratory failure, and kidney failure. DuKA utilizes pre-trained embeddings of medical codes and combines them using a modality-wise attention module and a medical concept-wise attention module to enhance interpretation. Three organ failure tasks are addressed using two datasets to verify the effectiveness of DuKA. Results: The proposed multi-modality DuKA model outperforms all reference and baseline models. The diagnosis history, particularly the presence of cachexia and previous organ failure, emerges as the most influential feature in organ failure prediction. Conclusions: DuKA offers competitive performance, straightforward model interpretations and flexibility in terms of input sources, as the input embeddings can be trained using different datasets and methods. Significance: DuKA is a lightweight model that innovatively uses dual attention in a hierarchical way to fuse diagnosis, procedure and medication information for organ failure predictions. It also enhances disease comprehension and supports personalized treatment.**

I. Introduction

Organ failure is the main cause of death in the Intensive Care Units (ICUs) [1], [50]. The heart, kidneys and the respiratory system are vital organs with high failure prevalence which leads to unplanned ICU admissions. Heart failure (HF) affects over 20 million people globally and has one-year mortality of around 20% [6], [46]; kidney failure (KF) has a similarly high one-year mortality rate and much higher incidence (over 50%) of patients who require dialysis. Moreover, KF patients also have prolonged hospital stays [23], [38]. Respiratory failure (RF) has the highest incidence rate in ICU and the in-hospital (short-term) mortality can rise 40% depending on the aetiology [21], [47]. Therefore, early identification of these vital organ failures not only has clinical significance but also alleviates national health expenditure burdens. However, current works mostly focus on single-organ failure prediction [3], [22], [34], [54]; little research is available on the multiple-organ failure states.

Electronic health record (EHR) systems store rich medical information of patients during their hospital admissions including medical histories, diagnoses, surgeries, etc. In particular, diagnosis information has been shown to have strong predictive power for diseases such as heart failure, mortality or readmission in multiple studies [9], [10], [22], [49], [53]. To improve the prediction accuracy, other information is added to the model input such as procedures [8], [48], [51], medications [52], [57] and demographics [22].

With such multi-modal high-complex data embedded in EHR, extracting informative representations for these medical concepts is a key for clinical tasks. Advances in representation learning methods for natural language processing have stimulated the development of models such as Word2Vec [35], GloVe [43], Transformers [55] and BERT [11], which have been successfully applied to clinical settings [9], [22], [49]. The learnt representations can be used directly as input features or in model initialisation for different kinds of downstream tasks.

Another critical part of these successes comes from the application of deep learning models. These models can further construct non-linear representations which are specifically tuned for the downstream tasks [4]. However, their lack of interpretability is the main limiting factor preventing widespread application of deep learning models to popularise in the clinical settings. Attention mechanisms emerged as a consequence [7], [39]. With attention integrated into a neural network, we are able to see interpretable results at the same time highly achieving models [9], [33], [48].

In this work, we incorporated the aforementioned modalities in EHR, the medical concepts of diagnoses, procedures and medications, as well as demographics to predict organ failures.

Z. Liu was supported by the Suzhou Industrial Park, China, Jiangsu Provincial Double Innovation Talent Programme. D. A. Clifton was supported by the Pandemic Sciences Institute at the University of Oxford; the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre (BRC); an NIHR Research Professorship; a Royal Academy of Engineering Research Chair; and the InnoHK Hong Kong Centre for Centre for Cerebro-cardiovascular Engineering (COCHE). (Corresponding author: Zhangdaihong Liu).

Z. Liu and D. A. Clifton are with the Department of Engineering Science, University of Oxford, Oxford, UK and the Oxford-Suzhou Centre for Advanced Research (OSCAR), Suzhou, China. (e-mails: Jessie.Liu@oxford-oscar.cn, david.clifton@eng.ox.ac.uk).

X. Wu is with the National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China (e-mail: wuxuan@smail.nju.edu.cn)

Y. Yang is with the School of Public Health, Shanghai Jiao Tong University School of Medicine, Shanghai, China (e-mail: emma002@sjtu.edu.cn).

We further designed a neural network model DuKA (DUal Keyless Attention) that facilitates the fusion of data at both the modality and concept levels. This model leverages attention mechanisms to learn the contributions of different modalities and concepts to the clinical tasks. DuKA employed attention modules in a hierarchical fashion that are designed specifically for modeling structured multi-modal EHR data. Notably, these attention modules enable straightforward interpretability of the model's predictions, allowing for ease of application ~~in clinical practice~~ across various levels of data granularity.

Most of the previous works require multiple historical hospital visits of patients for prediction which is demanding for data storage/collection. In our setting, we simplified the tasks to be one-time-step predictions, i.e. only information within one hospital visit is needed for prediction. This setting is more applicable for use in low/mid-income countries where health records systems are typically not connected between hospitals, making it difficult to track the full medical records of patients treated at different hospitals.

We validated our model on the MIMIC-IV [19] and eICU Collaborative Research Database [45] datasets with three organ failure tasks: a multi-class prediction task, predicting the organ failure type among organ failure patients, and two binary prediction task, predicting organ failure among essential hypertension patients and ICU patients. Being able to distinguish the three organ failure types requires a comprehensive background in the medical specialty of each organ failure and is clinically challenging. This is the motivation for designing the first task. Moreover, essential hypertension is highly prevalent in our population and the dataset. It is also regarded as a risk factor of the considered organ failures [18]. Therefore, the second task aims to identify the risk of essential hypertension patients developing organ failures. The third task validates the model on a different dataset and targets at ICU patients which is a group of patients with high probabilities to suffer organ failures, especially respiratory failure [28], [41].

In practical applications, clinical datasets often pose challenges due to their limited sample sizes and complex structures. These challenges can lead to the over-parameterization of large models or models that are not well-suited to the data. Furthermore, neural network models with intricate architectures often lack interpretability. In light of these issues, this study demonstrates how DuKA addresses these challenges in the context of three important clinical tasks. The contributions of our work can be summarized in three main aspects: 1) Introduction of DuKA: We propose DuKA as a dedicated model for modeling multi-modal EHR data. DuKA is designed to be a lightweight model that enables the fusion of medical concepts, modalities, and offers interpretability. 2) Dual-keyless attention incorporation: We are the first to incorporate keyless attention in a dual manner for the purpose of EHR data fusion. This innovative approach allows for the effective integration of information from various modalities in a cohesive manner. The resulting attention scores can aid clinical practice by indicating the information that holds greater importance for organ failure predictions. 3) Novelty in predicting vital organ failures: To the best of our knowledge, this work represents the first attempt at predicting multiple vital organ failures simultaneously by utilizing fused information from diagnoses, procedures, and medications.

II. Related Work

There are two key components in DuKA: the medical concept representation learning and interpretable multi-modal fusion using neural attention. The second component in fact has two fundamentals, multi-modal fusion and interpretable neural network.

Previous works have shown that simple multi-hot encoding of the medical codes may be inferior to the pre-trained dense vector embeddings due to the pre-trained embeddings' ability of capturing local/global information [26], [58]. Therefore, extracting informative representations from EHR data is vital for clinical tasks, therefore, has been extensively studied in recent research. One highly-cited early work used GloVe [43] to train diagnosis code embeddings (representations) and gained success in several downstream tasks including heart failure prediction [9]. GloVe is a context-free representation learning method that is particularly good at capturing global information since it uses global co-occurrence of codes to generate embeddings. In clinical settings, the co-occurrence information is a valuable source of information for learning medical code embeddings since relationships between diseases are complicated and the co-occurrence of diseases reveals the latent pattern to some extent. Therefore, GloVe has been widely-applied in recent studies [25], [58]. BERT [11] is another popular method for training medical concept embeddings due to its huge success in natural language processing. Studies such as [22], [49], [52] all adapted BERT for medical code pre-training and downstream tasks. BERT gained its popularity in clinical applications due to its strong ability to extract contextual information, allowing patterns hidden between medical codes/concepts to be utilized. Moreover, the high complexity of the model architecture brings superior performance to the tasks. Other methods such as Word2Vec [36] and ELMo [44] are also widely applied in learning concept/code representations. Comparatively, GloVe and BERT have the most competitive performance and can be considered as good representatives for the context-free and contextual model categories, respectively [14], [20], [58].

Neural network models have shown promising performances in tackling clinical tasks such as disease prediction/classification and achieved high accuracies. Since clinical applications are demanding for interpretability, attention mechanisms were invented for improving the interpretability of neural network models. Both [9] and [52] proposed graph-based attention models to pre-train medical code embeddings; [10] used recurrent neural network with attention to pre-train medical code embeddings. Another way to embed attention is to directly use it in predictive models. Models like the Transformers and BERT have in-house attention modules. Although BERT models can achieve high performance, having multiple attention heads at each layer requires integration approach for attention interpretation and thus, is not straightforward. Moreover, they work on sequential data and are computationally expensive. Notably, all attention mechanisms

applied in the aforementioned works require a key-value pair or key-query-value triplet. There are application scenarios where such requirements can not be satisfied.

Many previous studies used multi-modal EHR data as input for clinical prediction tasks. One early work [8] used diagnosis, procedure, and medication codes to predict diagnosis and/or medications, however, these codes were simply put together and treated equally. We lose the modality-specific information contained between these codes. [10] used diagnosis codes and treatment (medication/procedure) codes in a hierarchical fashion to predict heart failures. [52] used medication, diagnosis codes and their ontologies to predict medications. The more recent work [22] incorporated diagnosis and demographics together to predict diagnoses. Almost all works that utilized medical codes used pre-trained or randomly initialised embeddings to represent them. If embeddings from different modalities are pre-trained using different methods or trained separately, they cannot be fused directly since the embeddings learnt represent different latent spaces. When the embeddings from different modalities/data spaces were concatenated, which was a common choice in most of the multi-modal works, it was assumed that these modalities contribute equally to the task which is also a limited way of achieving fusion [17], [27]. Attention mechanism is a good remedy for multi-modal fusion and has been widely applied for this purpose in the areas of computer vision, natural language processing as well as biomedical engineering [17], [24], [29], [37], [59].

III. Method

A. Embedding Pre-train

Taking MIMIC dataset as an example, for each hospital visit, we extracted three sets of medical codes/concepts representing the diagnoses, procedures and medications that a patient acquired during a visit. In particular, for diagnosis and procedure, we used ICD-9-CM (International Classification of Diseases, Ninth Revision, Clinical Modification) codes; we used the medication names directly instead of any drug codes so that they can be easily matched to any coding system. Notably, the pre-trainings for diagnosis and procedures were performed over the whole MIMIC-IV cohort that used ICD-9 codes; the dataset used for pre-training medication embeddings was the whole MIMIC-IV cohort.

1) GloVe: For each of the three modalities, we further constructed a co-occurrence matrix based on each set of the medical codes/concepts separately. Taking diagnosis as an example, the training data was $\mathscr{D} = [\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_k]$, where $\mathbf{v}_i$ represents a hospital visit and k is the total number of hospital visits in the dataset. Moreover, $\mathbf{v}_i = [c_1, c_2, ..., c_d]$ where c_i represents a diagnosis code and d is the total number of codes occurred in that visit. If two codes occurred together within a visit, the value at the corresponding entry in the co-occurrence matrix got updated. Lastly, we applied GloVe [43] separately to the three co-occurrence matrices and set the embedding dimensionality as 128.

2) BERT: For each of the three modalities, the BERT model received patients' visiting sequences encoded in this modality as patient-level pre-training data. Specifically, the pre-training data was $\mathscr{D}_{\text{patient}} = [\mathbf{p}_1, \mathbf{p}_2, ..., \mathbf{p}_k]$, where $\mathbf{p}_i = [\mathbf{v}_{i,1}, ..., \mathbf{v}_{i,i_k}]$ represents the chronically ordered ith patient's visit sequence; $i \in [1, k]$, i_k is the length of his visit sequence, and each $\mathbf{v}_{i,j}$, $j \in [1, i_k]$ is a training-sample in GloVe described above. We adopted the structure of the model in Med-BERT [49] pre-trained with masked language model (MLM) objective. $\mathscr{D}_{\text{patient}}$ was augmented by adding '[SEP]' token between two visits and '[CLS]' token in the beginning of each $\mathbf{p}_i$. 15% of codes in each $\mathbf{p}_i$ sequence were masked, 80% of which were replaced with '[MASK]' token, 10% were replaced with random token except '[MASK]', and the rest part remained unchanged. Each patient's visit sequence was then embedded into two embeddings – code embedding and segment embedding. In particular, codes belonging to the same visit would have the same segment embedding, e.g. codes in $\mathbf{v}_{i1}$ were all embedded with the first segment embedding, for all $i \in [1, k]$. Through training, the MLM objective led the code embedding to learn the contextual information in each patient's visits and to predict the co-occurrence between these codes. The dimensionality of code embedding was also set as 128.

Similarly, we used the aforementioned methods to pre-train medical concept embeddings for the eICU dataset following the same processing pipeline. Notably, for procedure information, we selected the surgeries/operations from the 'treatment' table and used the procedure names directly rather than ICD codes [1].

B. Dual Keyless Attention Model

DuKA aims to model the multi-modal EHR data by utilizing medical concept/code embeddings that have been pre-trained using the methods discussed in the previous section. It then outputs the probability of a clinical event based on the specific task. In this study, the inputs consists of embeddings from three medical modalities (diagnosis, procedure, and medication) that occurred within a patient's hospital visit/ICU admission, and the output is the probability of experiencing organ failure during the patient's subsequent hospital visit or ICU admission. This scenario involves two levels of complexity. Firstly, within each modality, there are numerous medical codes, resulting in a large number of high-dimensional embeddings that need to be integrated. Secondly, the embeddings from different modalities cannot be integrated (e.g. by taking the average) directly since they are pre-trained in the context of the single modality and therefore not in a shared common space. To address the first level of complexity, the code-wise keyless attention mechanism is employed to fuse multiple code embeddings into a single representation that captures the information of the modality to which the codes belong. To tackle the second level of complexity, the modality-wise keyless attention mechanism is utilized to integrate all modality-level embeddings into a unified representation that captures cross-modality information.

As illustrated in Fig. 1, DuKA fuses multi-modality input and simultaneously offers interpretation at two different data levels, code/concept-level and modality-level. DuKA takes in

[1] https://eicu-crd.mit.edu/eicutables/treatment/

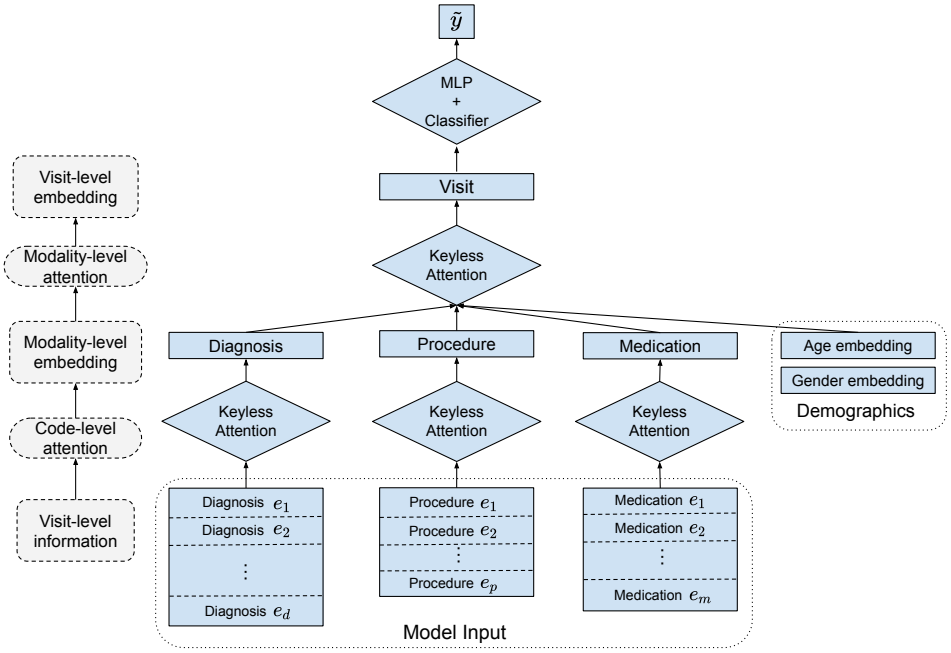


Fig. 1: Illustration for the Dual Keyless Attention (Duka) Model. The e in the Model Input box represents the pre-trained embeddings and its subscriptions d, p and m represent the number of diagnosis, procedure and medication codes in a visit, respectively. Notably, the age and gender embeddings are randomly initialised which follows the practices implemented in [22].

three sets of pre-trained embeddings that are matched with the visit-level information (the three sets of codes occurred during a visit). They are then fed into keyless attention modules separately to generate modality-level embedding and learn the code importance. Notably, since this study incorporates three different data modalities, the keyless attention mechanism is invoked three times at the code-level, once for each data modality. Secondly, the modality-level embeddings are further fed into a second attention module to generate visit-level embedding which is lastly used to perform the task. This completes the dual attentions. The keyless attention module handles missing modality/code by using masked attention, attending the non-empty modalities/codes only. Moreover, we can also choose to include patients' age and gender as predictors. Finally, the model input features are the pre-trained embeddings for diagnosis, procedure, medication and random initialised, trainable age and gender embeddings.

1) Keyless attention: We adopted a keyless attention mechanism in DuKA to fuse embeddings from different modalities and learn embedding importance. The original attention that was first proposed in LSTM (Long-short term memory) [2], [32] requires a key/anchor to calculate the attention scores. More recently, a popular attention mechanism was proposed in the Transformer work [55] which requires a key-query pair to learn attention scores. These different attention mechanisms are illustrated in Fig. 2. More detailed differences are explained in Appendix II.

Specifically, the keyless attention is calculated as follows: taking the code-level attention as an example, the attended output embedding which we denote as the modality-level embedding, z, is computed as

$$z = \sum_i \alpha_i e_i, \tag{1}$$

where e_i is the pre-trained code embedding within a modality; α_i is the attention score and calculated as

$$\alpha_i = \frac{\exp(h_i)}{\sum_j \exp(h_j)}. \tag{2}$$

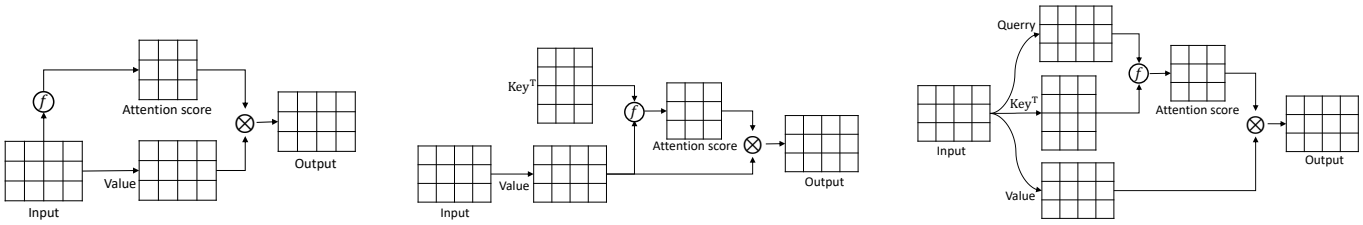
h_i is a function of e_i and has the same form as the paper that first proposed attention ([2]), a multi-layer perceptron (MLP) with tanh as the activation function, specifically,

$$h_i = f(e_i) = v_i^\top \tanh(w_i^\top e_i), \tag{3}$$

where w_i and v_i are the trainable weight vectors of the two hidden layers in the MLP. Note that function $f(\cdot)$ in Eqn. 2 now acts only on one object e_i, instead of a pair or a triplet which the usual attention mechanisms operate on.

Similarly, the second-level (modality-level) of attention is computed in the same way using Eqns. 2 and 3. The attended embedding z in Eqn. 2 is now the visit-level embedding (the top grey dotted box in Fig. 1), and the input e_i become the modality-level embeddings, i.e. the z output from the previous attentions.

We found similar usage of the keyless attention in the area of computer vision [29], [31]. In our use cases, complex models such as BERT that employs the key-value-query attention struggled to converge, potentially due to over-paramerization. Large and complex models typically involve a large number of parameters, which can make them challenging to train and optimize on smaller datasets. The key-value-query attention



(a) Keyless (b) Key-value (c) Key-query-value

Fig. 2: Illustration of the different attention mechanisms. (a) shows the keyless attention adopted in this work which learns the attention score by operating on the input data itself. (b) is a form of key-value attention in which the attention score is calculated by requiring a Key matrix. (c) illustrates the self-attention introduced in the Transformer model [55] where the attention score is obtained via the key-query relationship and then mapped to the Value matrix. Together, the key-query pair can be regarded as the 'key' in the attention mechanism.

mechanism, while powerful, may exacerbate this problem by introducing additional parameters and increasing model complexity. In contrast, keyless attention offers a more efficient alternative for modeling smaller clinical datasets. By eliminating the need for explicit keys, values, and queries, keyless attention reduces the parameter overhead. This streamlined approach simplifies the model architecture and improves its ability to generalize to limited data.

The computational complexity is detailed in Appendix III.

C. Ablation study

For ablation studies, we compared DuKA with the single-modality single-attention models. The model input is simply the pre-trained embeddings of diagnosis/procedure/medication. The model naturally becomes a single-attention model since it does not require modality-level attention. Notably, not every visit has all three modalities' information. We trained the single-modality models using data without missingness. Therefore, the sample size for the single-modality models is different.

D. Baseline models

We also tested several baseline models including random forest classifier (RFC), gradient boosting classifier (GBC), stochastic gradient descent classifier (SDGC) and one versus the rest classifier (OVRC). The model input is the concatenation of the averaged embeddings of each modality.

E. Model Assessments

To assess the above model, we split the respective dataset into training, test and validation sets. The test set separation is at the patient level to avoid leaking of intra-subject patterns: we split the patients into a training+validation cohort and a test cohort at the ratio of 0.8:0.2. The training+validation cohort is further unwrapped to visit-level samples and split into training and validation sets with the same ratio (0.8:0.2). Finally, the test cohort is unwrapped to visit-level samples to allow model testing. The class distributions between the three sets were ensured to be similar.

We assessed the model performance using the weighted and macro averages of precision, recall and F1-score of each class, area under receiver operating characteristic (AUROC) and confusion matrix. We ran each set of experiments 10 times to obtain the mean and standard deviation of the assessment measures. All results will be reported for the test set only, and all models were trained and tested on the same splits of data.

IV. MIMIC Task1: Multi-class Organ Failure Type Prediction

A. Data and Task Setting

We set the task to predict which one of the three organ failures will occur on the next visit. We further constrained that the next visit happens within six months. We labelled each visit based on the diagnosis code and the label can be one of the three organ failure types, HF, RF or KF. Notably, we did not consider patients with multiple organ failures. To simplify the application in the real world, we assumed that the most recent visit has the strongest impact on the next visit for organ failure patients, therefore, only information in the most recent visit was used to predict the organ failure label of the next visit. The task is thus visit-based and we further enlarged the dataset by unwrapping patients' visits. For example, a patient with three visits can construct two training samples: visit 1 to predict the label of visit 2 and visit 2 to predict the label of visit 3.

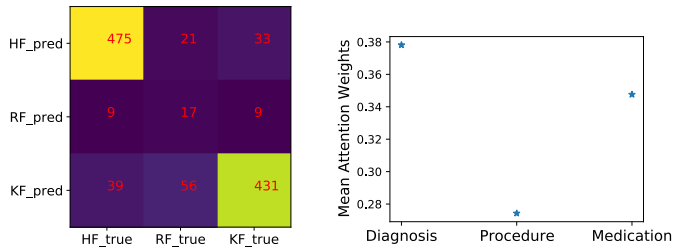
In MIMIC-IV, we selected hospital admissions whose diagnosis codes are stored using the ninth version of ICD (ICD-9). To identify patients with the three types of organ failure, we worked with clinicians and selected 28 ICD codes related to these three organ failures (a full list of ICD codes can be found in Appendix Table V). The data pre-processing pipeline is shown in Fig. 7a. For the purpose of prediction, we selected patients with at least two hospital visits and excluded patients with multiple organ failures.

The data summary is shown in Table. I and the pre-processing pipeline is shown Appendix Fig. 7. We denote these patients as the *target cohort*. We later found that if we process the data following the pipeline in Fig. 7a, every of the 8306 visits has an organ failure diagnosis, i.e. we have no negative samples. Label-wise, the dataset is very imbalanced with HF:RF:KF $\approx$ 15:1:13.

The model implementation details can be found in Appendix III.

TABLE I: Data summaries for the two datasets and three tasks. '#' represents the 'number of'.

	# patients	# visits	class ratio	# diagnosis	# procedure	# medication
MIMIC Task 1	2927	8306	HF:RF:KF ≈ 15:1:13	2958	576	1359
MIMIC Task 2	9486	22323	Negative:Positive ≈ 3:1	5196	1315	2381
eICU	2583	6658	Negative:Positive ≈ 1:4	508	388	1178



(a) Confusion Matrix (b) Modality-wise attention score

Fig. 3: The confusion matrix and modality-wise attention score for Task 1. These results are drawn from the DuKA model with the highest AUROC.

B. Results

We show the results without adding demographics to inputs in Table II, since adding demographics did not improve model performance in this task. We attach the results with demographics in Appendix Table VI. Table II also gives the comparison between GloVe and BERT embeddings. For all baseline models and DuKA, GloVe embeddings outperformed BERT embeddings based on the AUROCs. Therefore, BERT embeddings were not further tested in the ablation study. DuKA model with GloVe embedding as input gives the best results with the mean AUROC being 90.978 which is significantly better than the second best – the diagnosis single-modality single-attention model (p-value = 0.009 in one-tail t-test). Moreover, among all single-modality models, using procedure information on its own has the worst predictive power.

By investigating the code-level attention scores, we found that cachexia, endoscopic retrograde cholangiopancreatography (ERCP) and furosemide are the heaviest-loaded variables for diagnosis, procedure and medication, respectively. Cachexia is a complex syndrome that is associated with many severe diseases such as heart failure, chronic pulmonary and kidney diseases and cancer [13], [56]. ERCP is a procedure used to treat the bile ducts and main pancreatic duct, and furosemide is a common medication used to treat heart failure, liver or kidney diseases. We attach the code-wise attention scores in Appendix Fig. 8.

We extracted the DuKA model with the highest AUROC and show its interpretations as an example. The confusion matrix in Fig. 3(a) and the modality-wise attention is shown Fig. 3(b). Due to the small sample size in RF, it has the worst performance among the three organ failures. The modality rank of the attention scores is in line with the single-modality models' performances with diagnosis having the most predictive strength and procedure having the least.

V. MIMIC Task2: Organ Failure Prediction For Essential Hypertension Patients

A. Data and Task Setting

The second task we performed is a binary classification task – predicting organ failures among essential hypertension patients. We identified patients with essential hypertension if any of their diagnosis codes start with '401' (Essential Hypertension). The data selection was very similar to the first task apart from using the essential hypertension ICD rather than organ failure ICD codes (Appendix Fig. 7b). The prediction is still visit-based. We labelled a patient's visit as an organ failure instance if the visit contains any of the organ failure ICD codes that were used in Task 1 without specifying which kind of organ failure.

Table. I shows the data summary for this dataset and the pre-processing pipeline is shown Appendix Fig. 7. More specifically, the positive rate is about 25.9% – 5792 out of 23223 total visits are organ failure visits.

Same with Task 1, we fed both GloVe and BERT embeddings with and without trainable age and gender embeddings as input to the models. The model implementation details are shown in Appendix III.

B. Results

Notably, the baseline models yielded closer results for the settings with and without using demographic measures. Fig. 4 shows that apart from SVM, all other models have GloVe embeddings with demographics as the best input setting, and Logistic Regression with GloVe embeddings and demographics has the highest AUROC.

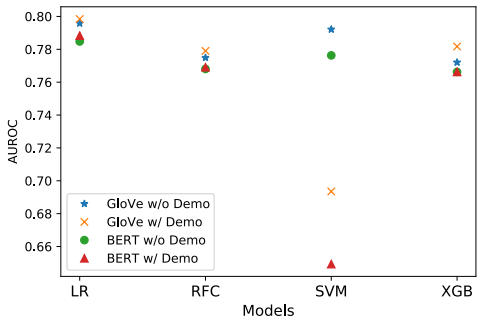


Fig. 4: The average AUROC for the four baseline models. We compared four sets of input, GloVe embeddings without demographics ('GloVe w/o Demo'), GloVe embeddings with demographics ('GloVe w/ Demo'), BERT embeddings without demographics ('BERT w/o Demo') and BERT embeddings with demographics ('BERT w/ Demo').

TABLE II: Organ failure type prediction results for the baseline (RFC, XGB, SGDC, and OVRC), DuKA and ablation study (the bottom block). The weighted scores are weighed by the number of labels in each class; the macro scores are the arithmetic means of the individual classes. 'Diagnoses', 'Procedure' and 'Medication' indicate the single-modality single-attention models in the ablation study. Due to the poor performance of BERT embeddings from the baseline and DuKA models, they were not tested for further ablation study. The AUROC for this multi-class setting is calculated by the 'one versus rest' approach due to the imbalance of the class labels. The numbers shown in the table are the average and standard deviation (in brackets) of the 10 random repetitions. '*': there is no randomness involved in XGB, therefore, no standard deviation is shown.

Model		Recall (%)		Precision (%)		F1-score (%)		AUROC (%)
		weighted (accuracy)	macro	weighted	macro	weighted	macro	
RFC	GloVe	82.61 (0.59)	57.79 (0.41)	81.79 (1.13)	73.45 (8.07)	81.42 (0.57)	57.76 (0.41)	86.06 (0.75)
	BERT	84.14 (0.43)	57.95 (0.30)	81.48 (0.42)	56.13 (0.29)	82.75 (0.43)	57.00 (0.29)	84.28 (0.81)
XGB*	GloVe	84.86	62.01	83.84	70.20	84.08	63.44	87.00
	BERT	85.78	59.08	83.22	57.32	84.44	58.16	83.38
SGDC	GloVe	72.20 (3.11)	65.81 (2.06)	85.16 (0.84)	61.38 (0.81)	82.21 (2.10)	57.81 (2.00)	85.98 (1.35)
	BERT	71.83 (2.05)	60.67 (2.20)	84.58 (1.04)	60.16 (0.83)	77.08 (1.52)	56.33 (1.23)	83.47 (1.16)
OVRC	GloVe	83.99 (1.69)	62.47 (2.93)	83.34 (1.95)	68.04 (4.85)	83.38 (1.81)	63.46 (3.16)	87.15 (1.85)
	BERT	86.20 (0.84)	61.24 (1.21)	84.81 (0.85)	65.77 (3.95)	85.28 (0.79)	61.68 (1.64)	84.38 (1.56)
DuKA	GloVe	89.56 (0.25)	67.94 (0.81)	85.95 (1.06)	74.94 (1.08)	87.47 (0.65)	69.15 (1.16)	**90.98 (0.56)**
	BERT	88.36 (0.22)	65.92 (0.47)	84.34 (0.81)	71.60 (1.00)	86.07 (0.50)	66.62 (0.72)	86.02 (1.12)
Diagnosis	GloVe	89.23 (0.23)	67.64 (1.07)	86.23 (1.16)	73.26 (1.07)	87.51 (0.75)	68.77 (1.38)	90.13 (0.80)
Procedure	GloVe	73.96 (0.52)	55.96 (0.32)	64.05 (1.05)	61.12 (1.04)	67.72 (0.74)	54.07 (0.73)	78.01 (1.07)
Medication	GloVe	80.00 (0.75)	59.42 (0.55)	74.01 (0.78)	66.58 (1.76)	76.43 (0.52)	59.39 (0.60)	85.83 (0.41)

For DuKA, unlike the baseline models, we found that without using demographic embeddings gave better performance (Table III). Therefore, they were not tested in the ablation study. In the ablation study, single-modality model using only diagnosis codes has shown competitive performance with DuKA, the second-best in AUROC. However, it is still significantly worse than DuKA (p-value = 0.009 in one-tail T-test).

1) attention scores interpretation: We extracted the attention scores for the two attention modules of DuKA. The mean attention scores for diagnosis, procedure and medication are 0.604, 0.223 and 0.173, respectively. We notice that the contribution of diagnosis is significantly higher than procedure and medication, and the order of contribution between procedure and medication has changed from Task 1.

For the code-level attention scores, we take diagnosis as an example. Fig. 5 shows the top 20 most-weighted diagnosis codes. The top code is 'unspecified essential hypertension' which is not surprising since the target cohort in this task is patients with essential hypertension and it is indicative towards organ failure. We also see many organ failure-related historical diagnoses such as 'congestive heart failure' and 'unspecified acute renal failure', which suggests patients with organ failure histories are more likely to develop organ failure again. Moreover, high-prevalence chronic diseases such as diabetes, hyperlipidemia and anaemia also appear at the top of the list which is suggested by literature to have correlations with HF and KF [5], [15], [16]. The attention scores also reveal 'tobacco use disorder' as a highly-weighted diagnosis, which is reported to be related to RF [12]. Interestingly, 'unspecified depressive disorder' appears on the list which indicate its associations with organ failures.

The attention scores for procedure and medication are listed in Appendix Fig. 9. Notably, furosemide is the most attended medication again which is the same in the Task 1.

We further separated the organ failure positive and negative patients and interpreted their attention scores. Fig. 6 presents

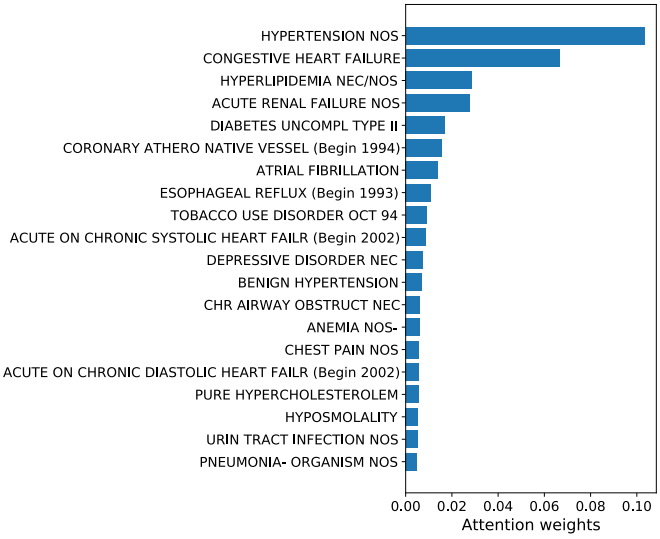


Fig. 5: Top 20 code-wise attention scores for diagnosis averaged across 10 repetitions of DuKA.

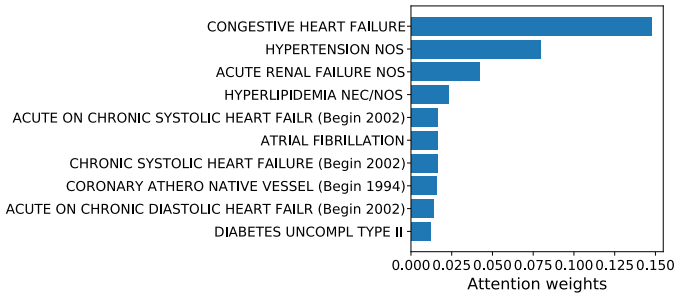
the top 10 most-weighted diagnosis codes for patients that developed organ failures (Fig. 6a) and did not develop organ failures (Fig. 6b). From these attention scores we can see that although the top diagnoses overlap largely between the two groups of patients, for patients that developed organ failures, having a congestive heart failure history is regarded as being most important by the model whereas for organ failure negative patients, congestive heart failure ranks sixth, weighing much lower than having hypertension.

We attach the same attention scores for procedure and medication in Appendix Figs. 10 and 11.

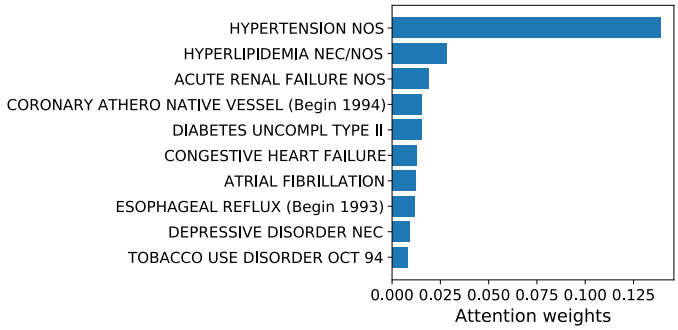
Notably, the subject-specific attention scores can be extracted from the model (illustrated in Appendix Figs. 12 and 13). These scores can assist with personalized treatment by guiding clinicians to prioritize specific diagnoses, procedures,

TABLE III: Task 2 results for DuKA (top block) and ablation study (bottom block). The weighted scores are weighed by the number of labels in each class; the macro scores are the arithmetic means of the individual classes. 'Diagnoses', 'Procedure' and 'Medication' indicate the single-modality single-attention models. Due to the inferior performance of BERT and demographic embeddings, they were not tested for the ablation study. The best model and AUROC are highlighted with bold font.

Model		Recall (%)		Precision (%)		F1-score (%)		AUROC (%)
		weighted (accuracy)	macro	weighted	macro	weighted	macro	
DuKA-BERT	w/ demo.	77.77 (0.64)	70.42 (1.28)	76.64 (1.44)	71.91 (0.60)	77.08 (1.17)	70.98 (1.09)	78.20 (0.64)
	w/o demo.	78.06 (0.73)	70.82 (1.61)	76.93 (1.56)	72.22 (0.58)	77.34 (1.21)	71.30 (1.71)	79.03 (0.35)
DuKA-GloVe	w/ demo.	78.50 (0.41)	70.92 (0.72)	76.90 (0.79)	73.18 (0.47)	77.50 (0.66)	71.76 (0.65)	79.49 (0.47)
	w/o demo.	78.30 (0.35)	70.72 (0.62)	76.75 (0.68)	72.90 (0.41)	77.34 (0.56)	71.54 (0.54)	**79.82** (0.28)
Diagnosis-GloVe	w/o demo.	78.20 (0.56)	70.56 (1.33)	76.46 (1.68)	72.70 (0.46)	77.07 (1.36)	71.27 (1.20)	79.49 (0.29)
Procedure-GloVe	w/o demo.	66.86 (0.70)	54.89 (0.98)	60.08 (3.98)	56.04 (1.04)	62.18 (3.40)	53.91 (2.12)	59.58 (1.06)
Medication-GloVe	w/o demo.	71.25 (0.47)	63.19 (0.72)	65.22 (2.01)	65.67 (0.74)	66.68 (1.86)	62.61 (1.44)	72.77 (0.43)



(a) Patients with organ failures



(b) Patients without organ failures

Fig. 6: Diagnosis attention score interpretation by different groups of patients. (a) shows the top 10 diagnosis codes for patients that develop organ failures; (b) are the top 10 diagnoses for patients without organ failure.

or medications that play a more crucial role in precipitating organ failure.

VI. DuKA validation on eICU database

Lastly, we tested the DuKA model on the eICU database. This dataset contains only ICU admissions over multiple centres in the US. Similar to the previous two tasks, we selected patients with one of the three organ failures based on the ICD-9 diagnostic codes. We further removed patients with only one ICU admissions. The data processing pipeline is shown in Fig. 7. The task is to predict whether a patient would experience one of the three organ failures in their next ICU admission, utilizing information collected from their previous ICU admission. The data summary is shown in Table I. Notably, more than 95% of the organ failures happened in ICU were respiratory failures. Given the nature of this dataset, the training unit was defined as the ICU admission rather than the hospital visit. The same model architecture and training pipeline used for the MIMIC datasets were maintained.

The results are shown in Table IV. Overall, DuKA shows satisfying performance on the independent eICU dataset in predicting organ failures in ICUs. The average modality-level attention weights across 10 random repetitions for diagnosis, procedure and medication are 0.54, 0.23, 0.24, respectively. The importance ranking between the three modalities remains similar to the two tasks on the MIMIC dataset.

VII. Discussions

The construction of DuKA takes two important factors into account. Firstly, DuKA is designed to fuse pre-trained medical code/concept embeddings originating from different modalities, which are trained separately. Leveraging pre-trained embeddings is a widely adopted approach due to their ability to provide meaningful data representations. By incorporating embeddings trained from diverse datasets or tasks, the model gains flexibility and facilitates transfer learning. Secondly, DuKA aims to maintain a simple model structure while maximizing interpretability. This is crucial for clinical applications where model interpretability holds significant value. By offering straightforward and simple feature importance, we prevent 'over-modeling' of relatively small clinical datasets by neural networks. Hence, instead of employing the multi-head module, we embed the keyless attention mechanism into DuKA. The resulting attention scores could aid the personalized treatment and specific task understanding in clinical practice. Additionally, we conducted an investigation into the attention scores from different repetitions and observed high stability. Overall, the proposed DuKA model addresses the challenges specific to modeling tabular EHR data.

However, we did not explore/optimise the form of the keyless attention. For example, one can try other forms such as taking the inner product of the feature itself. Moreover, we can use different attention forms/dimensions for the two attention modules. We are confident that these can improve the performance of DuKA and are interesting future directions to explore.

Another emphasis of this work was to compare two popular presentation learning methods and their pre-trained embeddings. We found that in almost all models we considered,

TABLE IV: DuKA validation results on the eICU dataset. Since the previous tasks showed that GloVe embeddings without demographic features have superior performance. This table shows the results for this setting only. The experiment was repeated for 10 random initialisations. The mean scores are shown for each assessment measure with the standard deviations in the brackets.

Recall (%)		Precision (%)		F1-score (%)		AUROC (%)
weighted (accuracy)	macro	weighted	macro	weighted	macro	
83.90 (2.01)	64.57 (0.92)	69.72 (3.67)	73.76 (2.48)	73.08 (3.09)	63.26 (1.86)	81.55 (0.68)

GloVe embeddings had better performance than BERT. We found similar results in another work that performed representation learning comparison [58]. The poor performance of BERT might be caused by that we used BERT to pre-train the embeddings with the masked language model only and did not fine-tune it using the downstream tasks. Besides, although this pre-train enables BERT to capture the contextual information within a patient's visit sequence, the global information about the medical codes of pathology is limited. By comparison, GloVe explicitly models the global co-occurrence information, which can give results that are more consistent with intuition.

We worked with clinicians and selected three types of vital organ failures to perform the tasks. We designed two prediction tasks, a multi-class organ failure type prediction and a binary organ failure prediction. These tasks only use information in one time step to predict the event in the next time point. We are aware that using less historical information may reduce the model's performance. However, this setting reduces the requirement for data acquisition and better suits the real-world scenarios in low to middle-income countries where no advanced EHR systems are in place or the EHR systems are not connected among hospitals and therefore, it is harder to track people's health history. Moreover, we also tried adding trainable age and gender embeddings using the same way with [22]. It is surprising to find that in most cases, adding them did not bring extra gain to the model performance, especially for DuKA. It may indicate that this way of incorporating demographic embeddings is not suitable for these tasks. It is also possible that for the tasks we conducted, age and gender are confounded with the diagnosis/procedure/medication information. We only considered two of the demographic measures. Future work can take more demographic/clinical features into consideration and apply more sophisticated approaches to handle them such as learning pre-trained embeddings.

In the conducted ablation studies, we observed that using single diagnosis modality as input achieved similar levels of performance in AUROC compared with the proposed DuKA model. However, through T-tests, we still identified significant improvements by employing DuKA. Moreover, one significant advantage offered by DuKA is that it allows clinicians to trace the contribution of variables from different sources of input (diagnosis/procedure/medication), which is meaningful in clinical practice. The modality-level attention scores offer valuable guidance to clinicians, encouraging them to prioritize diagnosis information when dealing with organ failure patients. This advice becomes especially crucial in time-constraining scenarios, such as admitting/treating patients in ICUs. The presented average code-level attention scores (e.g. Fig 5) could help with specific task understanding. The model can also generate subject-specific attention scores (Appendix Figa. 12 and 13), facilitating personalized treatment. This enables healthcare providers to proactively address the prioritized diagnostic, procedural, or medication requirements of each patient.

DuKA also allows incorporation of other data modalities which could potentially increase the performance gap from the single-modality models. This is a worthwhile direction for future investigations. One other limitation of this work is the selection of the target cohort. Although we worked closely with two clinicians to select the ICD codes, it is possible that some organ failure patients are omitted. This may cause biased data labelling and model results.

VIII. Conclusions

In this work, we introduce the Dual-Keyless Attention (DuKA) model for modeling tabular Electronic Health Record (EHR) data. The effectiveness of DuKA is demonstrated through its application on two datasets and three clinical tasks. The AUROCs received over these tasks range from 0.800 to 0.910. DuKA could further offer diagnostic, procedural and medication-related clinical interpretations that are relevant to the organ failures considered. Its ability to fuse embeddings from diverse EHR data modalities, provide interpretable results, and maintain simplicity in model architecture while maximizing interpretability showcases its potential value in clinical applications.

References

[1] B. Afessa, B. Green, I. Delke, and K. Koch. Systemic inflammatory response syndrome, organ failure, and outcome in critically ill obstetric patients treated in an icu. *Chest*, 120(4):1271–1277, 2001.

[2] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[3] M. Bartoletti, M. Giannella, L. Scudeller, S. Tedeschi, M. Rinaldi, L. Bussini, G. Fornaro, R. Pascale, L. Pancaldi, Z. Pasquini, et al. Development and validation of a prediction model for severe respiratory failure in hospitalized patients with sars-cov-2 infection: a multicentre cohort study (predi-co study). *Clinical Microbiology and Infection*, 26(11):1545–1553, 2020.

[4] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

[5] B. Bozkurt, D. Aguilar, A. Deswal, S. B. Dunbar, G. S. Francis, T. Horwich, M. Jessup, M. Kosiborod, A. M. Pritchett, K. Ramasubbu, et al. Contributory risk and management of comorbidities of hypertension, obesity, diabetes mellitus, hyperlipidemia, and metabolic syndrome in chronic heart failure: a scientific statement from the american heart association. *Circulation*, 134(23):e535–e578, 2016.

[6] I. Bytyçi and G. Bajraktari. Mortality in heart failure patients. *Anatolian journal of cardiology*, 15(1):63, 2015.

[7] S. Chaudhari, V. Mithal, G. Polatkan, and R. Ramanath. An attentive survey of attention models. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 12(5):1–32, 2021.

[8] E. Choi, M. T. Bahadori, A. Schuetz, W. F. Stewart, and J. Sun. Doctor ai: Predicting clinical events via recurrent neural networks. In *Machine learning for healthcare conference*, pages 301–318. PMLR, 2016.

[9] E. Choi, M. T. Bahadori, L. Song, W. F. Stewart, and J. Sun. Gram: graph-based attention model for healthcare representation learning. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 787–795, 2017.

[10] E. Choi, C. Xiao, W. Stewart, and J. Sun. Mime: Multilevel medical embedding of electronic health records for predictive healthcare. *Advances in neural information processing systems*, 31, 2018.

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[12] T. Dombernowsky, M. Ø. Kristensen, S. Rysgaard, L. L. Gluud, and S. Novovic. Risk factors for and impact of respiratory failure on mortality in the early phase of acute pancreatitis. *Pancreatology*, 16(5):756–760, 2016.

[13] N. Ebner, J. Springer, K. Kalantar-Zadeh, M. Lainscak, W. Doehner, S. D. Anker, and S. Von Haehling. Mechanism and novel therapeutic approaches to wasting in chronic disease. *Maturitas*, 75(3):199–206, 2013.

[14] E. Getzen, Y. Ruan, L. Ungar, and Q. Long. Mining for health: A comparison of word embedding methods for analysis of ehrs data. *medRxiv*, 2022.

[15] K. Gupta, R. Kalra, I. Rajapreyar, J. M. Joly, M. Pate, M. G. Cribbs, S. Ather, S. D. Prabhu, and N. S. Bajaj. Anemia, mortality, and hospitalizations in heart failure with a preserved ejection fraction (from the topcat trial). *The American journal of cardiology*, 125(9):1347–1354, 2020.

[16] R. M. Hanna, E. Streja, and K. Kalantar-Zadeh. Burden of anemia in chronic kidney disease: beyond erythropoietin. *Advances in therapy*, 38(1):52–75, 2021.

[17] C. Hori, T. Hori, T.-Y. Lee, Z. Zhang, B. Harsham, J. R. Hershey, T. K. Marks, and K. Sumi. Attention-based multimodal fusion for video description. In *Proceedings of the IEEE international conference on computer vision*, pages 4193–4202, 2017.

[18] A. M. Iqbal and S. F. Jamal. Essential hypertension. In *StatPearls [Internet]*. StatPearls Publishing, 2022.

[19] A. Johnson, L. Bulgarelli, T. Pollard, S. Horng, L. Celi, and R. Mark. Mimic-iv (version 1.0), 2020.

[20] J. M. Johnson and T. M. Khoshgoftaar. Medical provider embeddings for healthcare fraud detection. *SN Computer Science*, 2(4):1–15, 2021.

[21] S. W. Ketcham, Y. R. Sedhai, H. C. Miller, T. C. Bolig, A. Ludwig, D. Claar, J. I. McSparron, H. C. Prescott, M. W. Sjoding, et al. Causes and characteristics of death in patients with acute hypoxemic respiratory failure and acute respiratory distress syndrome: a retrospective cohort study. *Critical Care*, 24(1):1–9, 2020.

[22] Y. Li, S. Rao, J. R. A. Solares, A. Hassaine, R. Ramakrishnan, D. Canoy, Y. Zhu, K. Rahimi, and G. Salimi-Khorshidi. Behrt: transformer for electronic health records. *Scientific reports*, 10(1):1–12, 2020.

[23] O. Liangos, R. Wald, J. W. O'Bell, L. Price, B. J. Pereira, and B. L. Jaber. Epidemiology and outcomes of acute renal failure in hospitalized patients: a national survey. *Clinical journal of the American Society of Nephrology*, 1(1):43–51, 2006.

[24] F. Liu, B. Yang, C. You, X. Wu, S. Ge, Z. Liu, X. Sun, Y. Yang, and D. Clifton. Retrieve, reason, and refine: Generating accurate and faithful patient instructions. *Advances in Neural Information Processing Systems*, 35:18864–18877, 2022.

[25] W. Liu, C. Stansbury, K. Singh, A. M. Ryan, D. Sukul, E. Mahmoudi, A. Waljee, J. Zhu, and B. K. Nallamothu. Predicting 30-day hospital readmissions using artificial neural networks with medical code embedding. *PloS one*, 15(4):e0221606, 2020.

[26] Z. Liu, Y. Hu, G. Mertes, Y. Yang, and D. Clifton. Patient clustering and classification for vital organ failure using icd code with graph attention. *bioRxiv*, 2022.

[27] Z. Liu, Y. Hu, X. Wu, G. Mertes, Y. Yang, and D. A. Clifton. Patient clustering for vital organ failure using icd code with graph attention. *IEEE Transactions on Biomedical Engineering*, 2023.

[28] N. I. Lone and T. S. Walsh. Impact of intensive care unit organ failures on mortality during the five years after a critical illness. *American journal of respiratory and critical care medicine*, 186(7):640–647, 2012.

[29] X. Long, C. Gan, G. De Melo, X. Liu, Y. Li, F. Li, and S. Wen. Multimodal keyless attention fusion for video classification. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[30] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[31] D. Luo, H. Xu, and L. Carin. Interpretable icd code embeddings with self-and mutual-attention mechanisms. *arXiv preprint arXiv:1906.05492*, 2019.

[32] M.-T. Luong, H. Pham, and C. D. Manning. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*, 2015.

[33] F. Ma, R. Chitta, J. Zhou, Q. You, T. Sun, and 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*, pages 1903–1911, 2017.

[34] G. Maragatham and S. Devi. Lstm model for prediction of heart failure in big data. *Journal of medical systems*, 43(5):1–13, 2019.

[35] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[36] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013.

[37] S. Mohla, S. Pande, B. Banerjee, and S. Chaudhuri. Fusatnet: Dual attention based spectrospatial multimodal fusion network for hyperspectral and lidar classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 92–93, 2020.

[38] K. Nash, A. Hafeez, and S. Hou. Hospital-acquired renal insufficiency. *American Journal of Kidney Diseases*, 39(5):930–936, 2002.

[39] Z. Niu, G. Zhong, and H. Yu. A review on the attention mechanism of deep learning. *Neurocomputing*, 452:48–62, 2021.

[40] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[41] P. B. Pedersen, D. P. Henriksen, M. Brabrand, and A. T. Lassen. Prevalence of organ failure and mortality among patients in the emergency department: a population-based cohort study. *BMJ open*, 9(10):e032692, 2019.

[42] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

[43] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[44] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. Deep contextualized word representations. arxiv 2018. *arXiv preprint arXiv:1802.05365*, 12, 1802.

[45] T. J. Pollard, A. E. Johnson, J. D. Raffa, L. A. Celi, R. G. Mark, and O. Badawi. The eicu collaborative research database, a freely available multi-center database for critical care research. *Scientific data*, 5(1):1–13, 2018.

[46] P. Ponikowski, S. D. Anker, K. F. AlHabib, M. R. Cowie, T. L. Force, S. Hu, T. Jaarsma, H. Krum, V. Rastogi, L. E. Rohde, et al. Heart failure: preventing disease and death worldwide. *ESC heart failure*, 1(1):4–25, 2014.

[47] H. C. Prescott, M. W. Sjoding, K. M. Langa, T. J. Iwashyna, and D. F. McAuley. Late mortality after acute hypoxic respiratory failure. *Thorax*, 73(7):618–625, 2018.

[48] Z. Qiao, X. Wu, S. Ge, and W. Fan. Mnn: multimodal attentional neural networks for diagnosis prediction. *Extraction*, 1:A1, 2019.

[49] L. Rasmy, Y. Xiang, Z. Xie, C. Tao, and D. Zhi. Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ digital medicine*, 4(1):1–13, 2021.

[50] Y. Sakr, S. M. Lobo, R. P. Moreno, H. Gerlach, V. M. Ranieri, A. Michalopoulos, and J.-L. Vincent. Patterns and early evolution of organ failure in the intensive care unit and their relation to outcome. *Critical care*, 16(6):1–9, 2012.

[51] E. Scheurwegs, B. Cule, K. Luyckx, L. Luyten, and W. Daelemans. Selecting relevant features from the electronic health record for clinical code prediction. *Journal of biomedical informatics*, 74:92–103, 2017.

[52] J. Shang, T. Ma, C. Xiao, and J. Sun. Pre-training of graph augmented transformers for medication recommendation. *arXiv preprint arXiv:1906.00346*, 2019.

[53] Y. Si, J. Du, Z. Li, X. Jiang, T. Miller, F. Wang, W. J. Zheng, and K. Roberts. Deep representation learning of patient data from electronic

health records (ehr): A systematic review. *Journal of Biomedical Informatics*, 115:103671, 2021.

[54] S. Tekale, P. Shingavi, S. Wandhekar, and A. Chatorikar. Prediction of chronic kidney disease using machine learning algorithm. *International Journal of Advanced Research in Computer and Communication Engineering*, 7(10):92–96, 2018.

[55] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[56] S. Von Haehling, N. Ebner, M. R. Dos Santos, J. Springer, and S. D. Anker. Muscle wasting and cachexia in heart failure: mechanisms and therapies. *Nature Reviews Cardiology*, 14(6):323–341, 2017.

[57] S. Wang, P. Ren, Z. Chen, Z. Ren, J. Ma, and M. de Rijke. Order-free medicine combination prediction with graph convolutional reinforcement learning. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 1623–1632, 2019.

[58] X. Wu, Z. Liu, Y. Zhao, Y. Yang, and D. A. Clifton. A comparison of representation learning methods for medical concepts in mimic-iv. *medRxiv*, 2022.

[59] T. Zhou, S. Ruan, Y. Guo, and S. Canu. A multi-modality fusion network based on attention mechanism for brain tumor segmentation. In *2020 IEEE 17th international symposium on biomedical imaging (ISBI)*, pages 377–380. IEEE, 2020.