Cardiac Complication Risk Profiling for Cancer Survivors via Multi-View Multi-Task Learning

Thai-Hoang Pham[1,2], Changchang Yin[1,2], Laxmi Mehta[3], Xueru Zhang[1], Ping Zhang[1,2]
[1]Department of Computer Science and Engineering, The Ohio State University, USA
[2]Department of Biomedical Informatics, The Ohio State University, USA
[3]Department of Medicine, Division of Cardiology, The Ohio State University, USA
{pham.375,yin.371,mehta.149,zhang.12807,zhang.10631}@osu.edu

***Abstract*—Complication risk profiling is a key challenge in the healthcare domain due to the complex interaction between heterogeneous entities (e.g., visit, disease, medication) in clinical data. With the availability of real-world clinical data such as electronic health records and insurance claims, many deep learning methods are proposed for complication risk profiling. However, these existing methods face two open challenges. First, data heterogeneity relates to those methods leveraging clinical data from a single view only while the data can be considered from multiple views (e.g., sequence of clinical visits, set of clinical features). Second, generalized prediction relates to most of those methods focusing on single-task learning, whereas each complication onset is predicted independently, leading to suboptimal models. We propose a multi-view multi-task network (MuViTaNet) for predicting the onset of multiple complications to tackle these issues. In particular, MuViTaNet complements patient representation by using a multi-view encoder to effectively extract information by considering clinical data as both sequences of clinical visits and sets of clinical features. In addition, it leverages additional information from both related labeled and unlabeled datasets to generate more generalized representations by using a new multi-task learning scheme for making more accurate predictions. The experimental results show that MuViTaNet outperforms existing methods for profiling the development of cardiac complications in breast cancer survivors. Furthermore, thanks to its multi-view multi-task architecture, MuViTaNet also provides an effective mechanism for interpreting its predictions in multiple perspectives, thereby helping clinicians discover the underlying mechanism triggering the onset and for making better clinical treatments in real-world scenarios.**

***Index Terms*—multi-view, multi-task, complication risk profiling, attention, insurance claims, contrastive learning**

I. Introduction

Cardiovascular diseases are widely known as the leading causes of mortality in breast cancer survivors [1]–[4]. With the recent substantial improvement of breast cancer survival rates, predicting the onset of multiple cardiac complications has become a critical task for enhancing patients' life quality. It is also a key to cost-effective disease management and prevention. However, this task is highly challenging because of the complex interactions between heterogeneous clinical entities. Effectively capturing these interactions may lead to more precise prediction and treatment for cancer survivors.

Over the past few decades, the rapid growth of real-world clinical data such as electronic health record (EHR) and insurance claims makes them valuable data sources used in data-driven (e.g., deep learning) systems for clinical risk prediction, especially complication risk profiling [5]–[7]. As shown in Figure 1, this data includes heterogeneous clinical entities (e.g., visit, disease, medication) and can be considered from multiple views (i.e., sequence of visits, set of features). However, most existing studies consider each clinical outcome prediction separately and extract information in clinical data from a single view, thereby, making them not well-suited for complication risk profiling and raising two challenges.

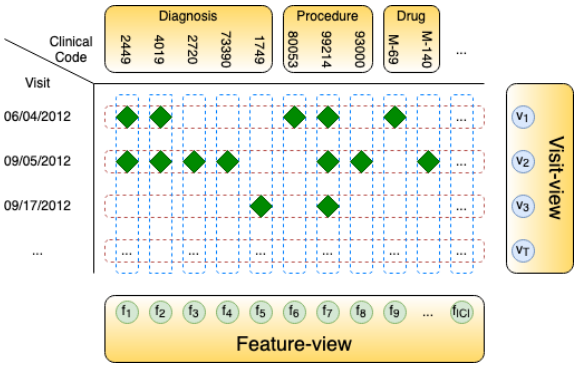


Fig. 1: Visit-view (sequence of clinical visits (rows)) and feature-view (set of clinical codes (columns)) of clinical data.

C1. Clinical data is highly complex due to its heterogeneous and hierarchical structure. Thus, encoding patient records from single-view cannot provide comprehensive representations of these patients, and thereby cannot achieve superior prediction performance. In particular, by considering patient records as sequences of visits, previous works only learn the dependencies among clinical visits but cannot explicitly capture dynamic patterns of clinical features and their interaction at the global (i.e., sequence) level.

C2. Treating each complication onset prediction independently can lead to suboptimal models, especially in limited datasets. This is because it cannot capture the dependencies among complications that are manifestations caused by their common underlying condition. Moreover, this approach cannot exploit meaningful clinical patterns from unlabeled data, which is much easier to collect and can be used to improve prediction performance when labeled data is limited.

To tackle the two aforementioned challenges, we propose a new neural network-based framework named Multi-View Multi-Task Network (MuViTaNet) for cardiac complication risk profiling. This proposed model consists of a multi-view encoder (dealing with **C1**) and a novel multi-task learning (MTL) scheme (dealing with **C2**). In particular, the **multi-view encoder** includes visit-view and feature-view encoders that capture information from clinical visits and features simultaneously. The visit-view encoder considers a patient record as the sequence of clinical visits and captures the temporal relation among visits by Gated Recurrent Unit (GRU) network. The feature-view encoder considers the patient record as the set of temporal medical features, and then leverages convolutional neural networks (CNN) to extract temporal patterns from these features separately. Then, the max-pooling operation is applied to extract the most significant signals from temporal sequences. The **MTL scheme** utilizes an attention mechanism to learn complication-specific representation from shared information generated by the multi-view encoder. This scheme allows MuViTaNet to exploit additional information from related complications and unlabeled data to generate more generalized representations for the patient, which enables more accurate predictions. Moreover, by leveraging the attention mechanism associated multi-view encoder, the proposed model provides an efficient way to interpret its predictions from multiple perspectives, thereby helping clinicians discover the underlying mechanism triggering the onset and making better clinical treatments. We demonstrate that the proposed model significantly outperforms current state-of-the-art approaches for complication risk profiling task using multiple datasets derived from the insurance claim database. In summary, our contributions include the following:

- We design a multi-view multi-task neural network architecture[1] (MuViTaNet) that accurately predicts multiple complication onsets and efficiently interprets its predictions.
- We develop a multi-view encoder to explicitly capture dependencies among clinical visits and clinical features from multiple views of clinical data.
- We also introduce a new MTL scheme that utilizes a complication-specific attention mechanism on top of the multi-view encoder to capture additional clinical information from related complications and unlabeled datasets.
- Finally, we conduct a comprehensive empirical study to demonstrate the effectiveness of MuViTaNet in terms of both prediction performance and interpretability compared to a wide range of previous approaches for cardiac complication risk profiling.

The remainder of the paper is organized as follows. Section II summarizes related works on clinical risk prediction in general and in particular, complication risk profiling. Section III describes the technical details of the proposed model (MuViTaNet). Section IV presents experimental results and discussions. Finally, Section V concludes the paper.

[1]Code is available at https://github.com/pth1993/MuViTaNet

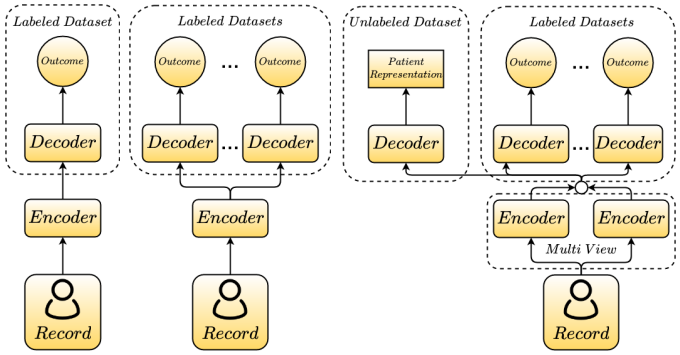
II. Related Works

In this section, we briefly review existing works related to our study including patient representation learning and MTL for clinical risk prediction, as well as complication risk profiling.

Patient representation learning. The abundance of real-world data in recent years creates an unprecedented opportunity to apply machine learning and data mining methods for clinical risk predictions. With the advancement of deep learning theory and the acceleration in computational technologies, neural network-based architectures can significantly improve prediction performance due to their ability to extract rich representations from data. Because of the temporal nature of clinical data, most existing methods rely on recurrent neural network architectures to learn patient representations, which are then used to make predictions for future clinical events (e.g., diagnosis, mortality, readmission, etc.) [5]–[9]. These works focused on designing attention mechanisms to capture dependencies among clinical visits [5], [8], [9] and time-aware mechanisms to incorporate temporal information [6], [10], [11] into patient representation for making better predictions. Nonetheless, these models cannot explicitly capture the relationships among clinical features. Instead of considering EHR data as sequences of clinical visits, Concare [12] treats the record as the set of clinical features and extracts dynamic patterns of these features separately. Then the predictions are made by aggregating representations of all clinical features. However, all the existing methods only extract information from a single view of clinical data which makes the learned patient representations suboptimal. In contrast, we propose a multi-view model for capturing information from multiple views of clinical data simultaneously.

Multi-task learning. Multi-task learning (MTL) has been used widely across many applications of machine learning and data mining. By sharing information among related tasks, the prediction model can generalize better. In healthcare domain, some existing works applied MTL techniques to leverage information from related tasks to improve model performance in clinical risk prediction. In particular, both classical machine learning [13]–[15] and deep learning models [16]–[18] are formulated as MTL frameworks and are applied on a wide range of healthcare applications including disease progression modeling [13], mortality prediction [16], disease onset prediction [17], and diagnosis classification [18].

Complication risk profiling. Mitigating the risk of complications is crucial for many disease management programs. Despite its importance, there have not been many existing methods designed for this task. Unlike a single clinical risk prediction task, complication risk profiling requires multiple predictions for onset of complications. Thus, capturing relationships among related complications is crucial to achieving good prediction performances. Some methods have been proposed to predict the onset of complications of some diseases and clinical procedures. For example, multi-task logistic regression has been used to predict complication risks for dia-



(a) (b) (c)

Fig. 2: General schemes for learning from clinical data. (a) single-view single-task learning, (b) single-view multi-task learning, (c) multi-view multi-task learning. Our proposed model belongs to multi-view multi-task learning with the multi-view encoder (i.e., visit-view and feature-view) and the task-specific attention mechanisms and decoders for both labeled and unlabeled datasets.

betes care [14], [19]. Besides linear models, the deep learning method is also used to predict complications of this chronic disease [20] but this work considers each complication independently. For breast cancer survivors, relationships between cardiac complications and cancer were also investigated [3], [4], [21] to show the correlation between these two diseases.

III. Methodology

In this section, we first give brief introduction about patient records, complication risk profiling task and the corresponding notations. Then, we present our proposed model MuViTaNet.

A. Definitions and Basic Notations

Definitions and notations used in this study are shown in the following paragraphs and are summarized in Table I.

Patient Record. The heterogeneous and hierarchical structure of a patient record is defined as follows.

- **Definition 1 (Clinical Code).** $\boldsymbol{C} = \{c_1, c_2, \cdots, c_{|\boldsymbol{C}|}\}$ is the set of unique clinical codes including diagnosis, procedure, and medication codes with $|\boldsymbol{C}|$ is the number of these unique codes. Each code c_i can be represented by binary vector $\boldsymbol{x}_i \in \{0,1\}^{|\boldsymbol{C}|}$ where i^{th} element of this vector is 1 and other elements are 0.
- **Definition 2 (Clinical Visit).** An visit is a hospital stay from admission to discharge. Each visit $\boldsymbol{v}_j$ is a tuple of $(\boldsymbol{c}_j,\ t_j)$ where $\boldsymbol{c}_j = \{c_{j_1}, c_{j_2}, \cdots, c_{j_{|\boldsymbol{c}_j|}}\} \in \boldsymbol{C}^{|\boldsymbol{c}_j|}$ with set of indexes $\{j_1, \cdots, j_{|\boldsymbol{c}_j|}\} \in \{1, 2, \cdots, |\boldsymbol{C}|\}$ and t_j is the timestamp of the visit. $\boldsymbol{c}_j$ can be represented by binary vector $\boldsymbol{V}_j \in \{0,1\}^{|\boldsymbol{C}|}$ where the i^{th} element is 1 if $\boldsymbol{c}_j$ contains the code c_i. Besides vector representation, $\boldsymbol{c}_j$ can also be expressed as matrix $\boldsymbol{X}_j \in \{0,1\}^{|\boldsymbol{c}_j| \times |\boldsymbol{C}|}$ where i^{th} row of this matrix is the binary vector $\boldsymbol{x}_{j_i} \in \{0,1\}^{|\boldsymbol{C}|}$ of code c_{j_i}.
- **Definition 3: (Patient Record).** The patient record $\boldsymbol{P}$ is a sequence of visits $[\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_T]$ where T is the number of visits. Like clinical visit representation, $\boldsymbol{P}$ can be represented at the two different granularities. At visit-level, $\boldsymbol{P}$ can be represented as a binary matrix $\boldsymbol{X}_{visit} \in \{0,1\}^{T \times |\boldsymbol{C}|}$ where j^{th} row of this matrix is binary vector $\boldsymbol{V}_j$ of visit $\boldsymbol{v}_j$. At feature-level, $\boldsymbol{P}$ can be represented as the sequence of matrices $\boldsymbol{X}_{feature} = [\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_T]$.
- **Definition 4: (Demographic Information).** Besides clinical information, a patient record can have demographic information about the patient such as age, gender, region, etc. It can be represented by binary vector $\boldsymbol{d}_{demo} \in \{0,1\}^{d_{demo}}$, where d_{demo} is the number of demographic attributes.

TABLE I: Notation definition

Notation	Description
$\boldsymbol{C}$	Set of clinical codes/features
$\boldsymbol{P}$	A patient record
c_i	i^{th} clinical codes in set C
$\boldsymbol{x}_i \in \{0,1\}^{\|\boldsymbol{C}\|}$	vector representation of code c_i
$\boldsymbol{v}_j$	j^{th} clinical visit in $\boldsymbol{P}$
$\boldsymbol{c}_j$	set of clinical codes in visit $\boldsymbol{v}_j$
t_j	timestamp of visit $\boldsymbol{v}_j$
$\boldsymbol{V}_j \in \{0,1\}^{\|\boldsymbol{C}\|}$	vector representation of visit $\boldsymbol{v}_j$
$\boldsymbol{X}_j \in \{0,1\}^{\|\boldsymbol{c}_i\| \times \|\boldsymbol{C}\|}$	matrix representation of visit $\boldsymbol{v}_j$
$\boldsymbol{X}_{visit} \in \{0,1\}^{T \times \|\boldsymbol{C}\|}$	visit-level representation of $\boldsymbol{P}$
$\boldsymbol{X}_{feature} \in T \times (\{0,1\}^{\|\boldsymbol{c}_i\| \times \|\boldsymbol{C}\|})$	feature-level representation of $\boldsymbol{P}$
$\boldsymbol{d}_{demo}$	vector representation of demographics
$\widehat{\boldsymbol{\alpha}}_j \in \mathbb{R}^{\|\boldsymbol{c}_j\|}$	attention weights of codes in visit $\boldsymbol{v}_j$
$\widehat{\boldsymbol{\beta}}_j \in \mathbb{R}^{\|\boldsymbol{C}\|}$	task-specific attention weights for features
$\widehat{\boldsymbol{\gamma}}_j \in \mathbb{R}^{T}$	task-specific attention weights for visits
$\boldsymbol{\delta}_j \in \mathbb{R}^{d}$	temporal encoding vector of visit $\boldsymbol{v}_j$
$\boldsymbol{H}^v \in \mathbb{R}^{T \times 2d}$	representation learned by visit-view encoder
$\boldsymbol{h}^* \in \mathbb{R}^{2d}$	patient representation
$\boldsymbol{H}^f \in \mathbb{R}^{\|\boldsymbol{C}\| \times 4d}$	representation learned by feature-view encoder
$\boldsymbol{g}_k^v \in \mathbb{R}^{2d}$	visit-view task-specific representation for k^{th} task
$\boldsymbol{g}_k^f \in \mathbb{R}^{4d}$	feature-view task-specific representation for k^{th} task
$\boldsymbol{o}_k \in \mathbb{R}^{8d}$	task-specific representation for k^{th} task
y_k	ground-truth output for k^{th} task
$\widehat{y}_k$	predicted output for k^{th} task

Clinical Risk Profiling. The aim of this task is to find a set of functions $\boldsymbol{F} = \{F_1, F_2, \cdots, F_N\}$ that predicts the onset of complications $\boldsymbol{Y} \in \mathbb{R}^N$ from patient record $\boldsymbol{P}$, where N is the number of complications. In MTL setting, $F_1, F_2, \cdots, F_N$ generally have some shared parameters to learn shared information from related tasks for better predictions.

B. Proposed Model

Overview Architecture. This section presents our proposed multi-view multi-task network (MuViTaNet) for predicting onset of multiple complications from patient records. MuViTaNet is designed to explicitly capture the dependencies among clinical visits and clinical features from patient records. It also leverages additional information from both related labeled and unlabeled data to achieve accurate predictions and efficient interpretation. In particular, MuViTaNet consists of four main components as follows. (1) Feature-view Encoder. This component considers a patient record as a set of temporal clinical features and then encodes information of each feature separately. (2) Visit-view Encoder. This component formulates a patient record as a sequence of visits and then learns a representation for each visit in the sequential context. Specifically, this component is designed as a hierarchical model that exploits patient records in the two levels, includ-

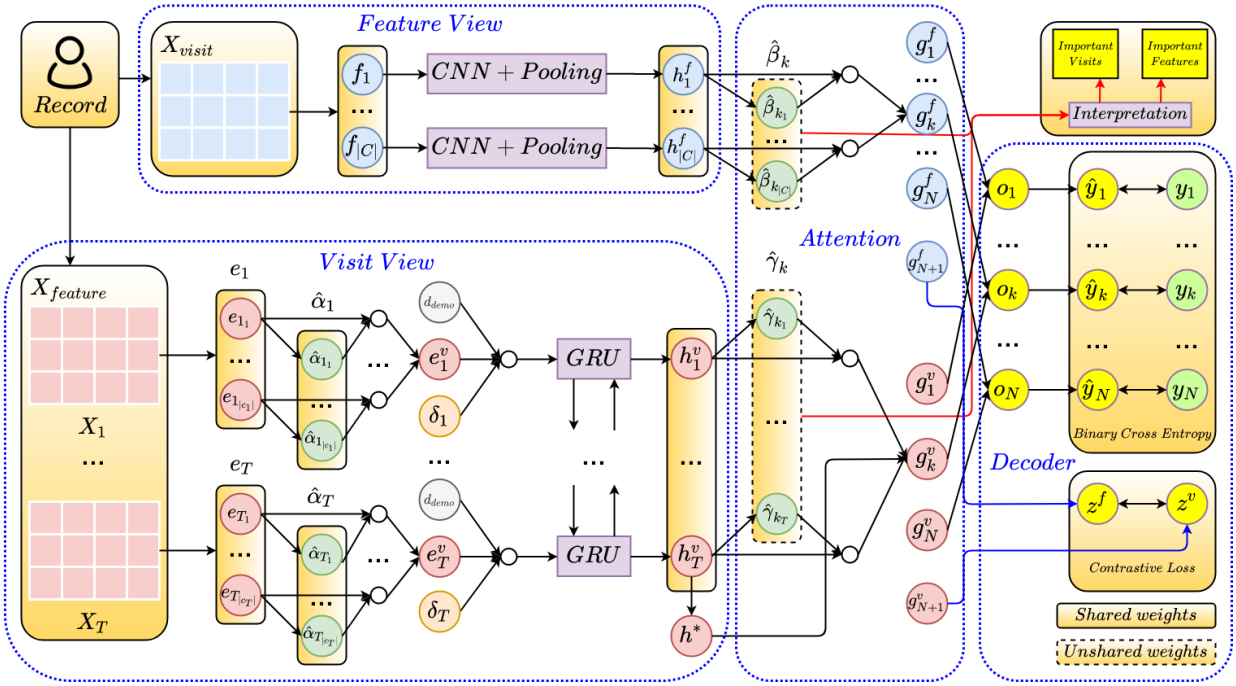


Fig. 3: The overall architecture of MuViTaNet. The proposed framework consists of four main components: feature-view encoder, visit-view encoder, task-specific attention, and task-specific decoder. Given a patient record, MuViTaNet first extracts information from clinical visits and features by looking at the record in two different ways: sequence of clinical visits and set of clinical features. Then the shared representation learned by these two encoders is put into the task-specific attention to learn the task-specific representation. Finally, the clinical predictions are generated by the task-specific decoders. Note that the figure only shows the task-specific attention for one prediction task for simplicity.

ing feature-level and visit-level. (3) Task-specific Attention. After learning the shared representation from feature-view and visit-view encoders, an attention mechanism is employed to extract task-specific representation for each task from the shared representation. (4) Task-specific Decoder. The task-specific representations are fed into the corresponding task-specific decoders to predict clinical outcomes for patients in complication datasets and to project representations to unit hypersphere for patients in unlabeled dataset. Figure 3 shows the overview architecture of MuViTaNet and technical details of its components are presented as follows.

Feature-view Encoder. This component treats patient data as a set $\boldsymbol{C}$ of clinical codes which are represented by the set of temporal sequences (i.e., columns of matrix $\boldsymbol{X}_{visit} \in \{0,1\}^{T \times |\boldsymbol{C}|}$). In particular, given clinical code c_i, its temporal data can be represented by a binary vector $\boldsymbol{f}_i \in \{0,1\}^T$ which is i^{th} column of $\boldsymbol{X}_{visit}$. Then, one-dimensional convolutional neural networks (Conv1d) and max-pooling (MaxPool) operation are employed to extract temporal patterns from each clinical code separately. In particular, Conv1d with kernel size k (i.e., $k = 3$ in our setting) takes as inputs the sub-sequences of length k from vector $\boldsymbol{f}_i$ to learn the representation of code c_i as follows.

$$\boldsymbol{H}_i^f = \text{Conv1d}(\boldsymbol{f}_i) \tag{1}$$

where $\boldsymbol{H}_i^f \in \mathbb{R}^{4d \times T}$ are the output of Conv1d and $4d$ is the number of filters used in convolution operations. Next, the row-wise max-pooling is applied to $\boldsymbol{H}_i^f$ to generate vector representation for clinical code c_i.

$$\boldsymbol{h}_i^f = \text{MaxPool}(\boldsymbol{H}_i^f) \tag{2}$$

Note that the weights of Conv1d are not shared between clinical codes. The output of feature-view encoder is matrix $\boldsymbol{H}^f = [\boldsymbol{h}_1^f, \boldsymbol{h}_2^f, \cdots, \boldsymbol{h}_{|\boldsymbol{C}|}^f] \in \mathbb{R}^{|\boldsymbol{C}| \times 4d}$.

Visit-view Encoder. This component formulates patient data as a sequence of visits in which each visit can be seen as a set of clinical codes. Due to the hierarchical characteristic of this data structure, the visit-view encoder is also designed hierarchically to capture information at different levels. Given visit $\boldsymbol{v}_j$, we represent this visit by matrix $\boldsymbol{X}_j \in \{0,1\}^{|\boldsymbol{c}_j| \times |\boldsymbol{C}|}$ which is j^{th} element of the sequence $\boldsymbol{X}_{feature}$. Because different clinical codes associated with the same visit can have disparate impacts, instead of treating these clinical codes uniformly when aggregating them to represent the visit, the location attention mechanism is employed to learn the contributions of these clinical codes to their visit representation. In particular, given a binary representation $\boldsymbol{x}_{j_i} \in \boldsymbol{X}_j$ of code c_{j_i}, 1-layer feed-forward neural network is applied to learn the dense representation from sparse vector of this clinical code as follows.

$$\boldsymbol{e}_{j_i} = \text{FFNN}_1(\boldsymbol{x}_{j_i}) = \text{ReLU}(\boldsymbol{W}_1 \boldsymbol{x}_{j_i} + \boldsymbol{b}_1) \tag{3}$$

Algorithm 1: Training procedure for MuViTaNet

Input: Datasets $\{\boldsymbol{D}_k\}_{k=1}^{N+1}$ (N labeled and 1 unlabeled datasets), set of clinical codes $\boldsymbol{C}$, batch sizes n_s, n_u

Output: Trained model parameters $\boldsymbol{\theta} = \{\boldsymbol{\theta}^{shared}, \{\boldsymbol{\theta}_k^{task-specific}\}_{k=1}^{N}\}$

Randomly initialize $\boldsymbol{\theta}$;
Calculate sampling rate for each dataset $\lambda_k = \frac{|\boldsymbol{D}_k|/n_k}{\sum_{k'=1}^{N}|\boldsymbol{D}_{k'}|/n_{k'}}$ ($n_k = n_u$ if $k = N+1$, $n_k = n_s$ otherwise);
for *epoch* $= 1$ *to* E **do**
  repeat
    Select dataset $\boldsymbol{D}_k \sim \boldsymbol{\lambda}$;
    Initialize loss $L_k = 0$;
    Select sample batch $\boldsymbol{b}$ from dataset $\boldsymbol{D}_k$;
    for *patient* $\boldsymbol{P}_i$ *in batch* $\boldsymbol{b}$ **do**
      $(\boldsymbol{X}_{feature}, \boldsymbol{X}_{visit}) = \boldsymbol{P}_i$;
      Obtain feature-view representation $\boldsymbol{H}^f$ from $\boldsymbol{X}_{visit}$ using Eq. (1), (2);
      Obtain visit-view representation $\boldsymbol{H}^v$ and patient representation $\boldsymbol{h}^*$ from $\boldsymbol{X}_{feature}$ using Eq. (3)-(11);
      Calculate task-specific attention weights $\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}}$ from $\boldsymbol{H}^f$, $\boldsymbol{H}^v$ using Eq. (12);
      Obtain task-specific representations using Eq. (13);
      if $k \in \{1, \cdots, N\}$ **then**
        Calculate prediction $\hat{y}_{k_i}$ using Eq. (14);
        Calculate BCE loss L_{k_i} using Eq. (16);
      else
        Project multi-view representations to unit hypersphere using Eq. (15);
        Calculate CL loss L_{k_i} using Eq. (17);
      $L_k = L_k + L_{k_i}$;
    end
    Update parameters $\boldsymbol{\theta}$ using gradient of L_k;
    $\boldsymbol{D}_k = \boldsymbol{D}_k \setminus \boldsymbol{b}$;
  until $\{\boldsymbol{D}_k\}_{k=1}^{N+1} == \varnothing$;
end

where $\boldsymbol{W}_1 \in \mathbb{R}^{d \times |\boldsymbol{C}|}$ is the learned weight matrix of clinical codes, $\boldsymbol{b}_1 \in \mathbb{R}^d$ is the bias vector, and ReLU is rectified linear unit activation function. Then the 2-layer feed-forward neural network FFNN_2 with Tanh activation function is used to generate the attention score α_{j_i} for this clinical code as follows.

$$\alpha_{j_i} = \mathrm{FFNN}_2(\boldsymbol{e}_{j_i}) \tag{4}$$

The attention vector $\boldsymbol{\alpha}_j = [\alpha_{j_1}, \alpha_{j_2}, \cdots, \alpha_{j_{|\boldsymbol{c}_j|}}]$ which represents the contributions of clinical codes in visit $\boldsymbol{v}_j$ is fed into the softmax layer to get the normalized vector $\widehat{\boldsymbol{\alpha}}_j = [\widehat{\alpha}_{j_1}, \widehat{\alpha}_{j_2}, \cdots, \widehat{\alpha}_{j_{|\boldsymbol{c}_j|}}] \in \mathbb{R}^{|\boldsymbol{c}_j|}$.

$$\widehat{\boldsymbol{\alpha}}_j = \mathrm{Softmax}(\boldsymbol{\alpha}_j) \tag{5}$$

Then, the representation of visit $\boldsymbol{v}_j$ are computed as the weighted average of its clinical codes.

$$\boldsymbol{e}_j^v = (\widehat{\boldsymbol{\alpha}}_j)^T \boldsymbol{e}_j \tag{6}$$

where $\boldsymbol{e}_j = [\boldsymbol{e}_{j_1}, \boldsymbol{e}_{j_2}, \cdots, \boldsymbol{e}_{j_{|\boldsymbol{c}_j|}}] \in \mathbb{R}^{|\boldsymbol{c}_j| \times d}$ denotes the j^{th} visit's representation. To generate personalized representation for each visit, demographic information including age and region is incorporated into every clinical visit as follows.

$$\ddot{\boldsymbol{e}}_j^v = \boldsymbol{W}_2(\mathrm{Concat}(\boldsymbol{e}_j^v, \boldsymbol{d}_{demo})) \tag{7}$$

where Concat is the concatenation operation and $\boldsymbol{W}_2 \in \mathbb{R}^{(d+d_{demo}) \times d}$ is the weight matrix mapping concatenated vectors to the original embedding space. Besides clinical codes, each visit is also associated with its timestamp. In order to capture the elapsed time between visits, we add the temporal encoding vector to each visit as follows.

$$\widehat{\boldsymbol{e}}_j^v = \ddot{\boldsymbol{e}}_j^v + \boldsymbol{\delta}_j \tag{8}$$

where $\boldsymbol{\delta}_j \in \mathbb{R}^d$ is the temporal encoding vector whose design is inspired by the positional encoding used in Transformer architecture [22]. In particular, it is computed by trigonometric functions as follows.

$$\begin{aligned} \delta_{j,2t} &= \sin\left(\frac{t_T - t_j}{10000^{2t/d}}\right) \\ \delta_{j,2t+1} &= \cos\left(\frac{t_T - t_j}{10000^{2t/d}}\right) \end{aligned} \tag{9}$$

where $0 \leq 2t < d-1$. From Equation (9), we can see that temporal embedding encodes similar time intervals into similar vectors in embedding space.

To generate the sequential representations for visits in the sequential context, we put the independent representations for visits learned from previous steps into the bidirectional GRU layer. Specifically, the sequential representation for these visits is computed as follows.

$$\begin{aligned} \overrightarrow{\mathbf{h}_j} &= \mathrm{GRU}(\widehat{\mathbf{e}}_j^v, \overrightarrow{\mathbf{h}_{j-1}}) \\ \overleftarrow{\mathbf{h}_j} &= \mathrm{GRU}(\widehat{\mathbf{e}}_j^v, \overleftarrow{\mathbf{h}_{j+1}}) \\ \mathbf{h}_j^v &= \mathrm{Concat}(\overrightarrow{\mathbf{h}_j}, \overleftarrow{\mathbf{h}_j}) \end{aligned} \tag{10}$$

where $\boldsymbol{h}_j^v \in \mathbb{R}^{2d}$. Then, the patient representation is computed based on the last visit in the visit sequence.

$$\boldsymbol{h}^* = \mathrm{FFNN}_3(\boldsymbol{h}_T^v) \tag{11}$$

In summary, the outputs of the visit-view encoder include the sequential representations of clinical visits $\boldsymbol{H}^v = [\boldsymbol{h}_1^v, \boldsymbol{h}_2^v, \cdots, \boldsymbol{h}_T^v] \in \mathbb{R}^{T \times 2d}$ and the patient representation $\boldsymbol{h}^* \in \mathbb{R}^{2d}$.

Task-specific Attention. Given the shared representations generated by feature-view and visit-view encoders, attention mechanisms are employed to generate the task-specific representations for the patient. Specifically, the attention weights of clinical features and visits for k^{th} task are computed as follows.

$$
\begin{aligned}
\beta_{k_i} &= \text{FFNN}_4^k(\boldsymbol{h}_i^f) \\
\gamma_{k_j} &= \text{FFNN}_5^k(\boldsymbol{h}_j^v) \\
\widehat{\boldsymbol{\beta}}_k &= \text{Softmax}([\beta_{k_1}, \beta_{k_2}, \cdots, \beta_{k_{|\boldsymbol{C}^l|}}]) \\
\widehat{\boldsymbol{\gamma}}_k &= \text{Softmax}([\gamma_{k_1}, \gamma_{k_2}, \cdots, \gamma_{k_T}])
\end{aligned} \tag{12}
$$

where $\text{FFNN}_4^k, \text{FFNN}_5^k$ are 2-layer feed-forward neural networks with Tanh activation function that compute the weights of clinical features and visits from their representations. Then, we obtain the task-specific representation $\boldsymbol{o}_k \in \mathbb{R}^{8d}$ for k^{th} task as follows.

$$
\begin{aligned}
\boldsymbol{g}_k^f &= (\widehat{\boldsymbol{\beta}}_k)^T \boldsymbol{H}^f \\
\boldsymbol{g}_k^v &= (\widehat{\boldsymbol{\gamma}}_k)^T \boldsymbol{H}^v \\
\boldsymbol{o}_k &= \text{Concat}(\boldsymbol{g}_k^f, \boldsymbol{g}_k^v, \boldsymbol{h}^*)
\end{aligned} \tag{13}
$$

Task-specific Decoder. For a patient in labeled dataset (i.e., complication dataset), the 2-layer feed forward neural network with Sigmoid activation function at the last layer is employed to predict the probability of complication onset for this patient.

$$
\hat{y}_k = \text{FFNN}_6^k(\boldsymbol{o}_k), \quad k \in \{1, \cdots, N\} \tag{14}
$$

For a patient in unlabeled dataset, the 2-layer feed forward neural network with normalization operation (Norm) is used to project the feature-view and visit-view representations of this patient on the unit hypersphere.

$$
\begin{aligned}
\boldsymbol{z}^f &= \text{Norm}(\text{FFNN}_6^k(\boldsymbol{g}_k^f)), \quad k = N+1 \\
\boldsymbol{z}^v &= \text{Norm}(\text{FFNN}_6^k(\text{Concat}(\boldsymbol{g}_k^v, \boldsymbol{h}^*)))
\end{aligned} \tag{15}
$$

Optimization. To train MuViTaNet in MTL setting, we follow the alternating training strategy [23] in which each task is selected randomly and then is optimized for a fixed number of parameter updates before switching to other tasks. In our setting, different tasks have datasets of different sizes, so we select a task to optimize with probability $\lambda_k = \frac{|\boldsymbol{D}_k| \backslash n_k}{\sum_{k'=1}^{N+1} |\boldsymbol{D}_{k'}| \backslash n_{k'}}$, where $\boldsymbol{D}_k$ and n_k are the dataset and batch size for k^{th} task, and N is the number of complication datasets.

For labeled datasets, the binary cross-entropy (BCE) loss function is used to optimize the prediction based on ground-truth labels. Specifically, for k^{th} task with dataset $\boldsymbol{D}_k$, the loss function for this task is computed as follows.

$$
L_L^k = -\frac{1}{|D_k|} \sum_{i=1}^{|D_k|} \Big(y_{k_i} \log(\widehat{y}_{k_i}) + (1 - y_{k_i}) \log(1 - \widehat{y}_{k_i}) \Big) \tag{16}
$$

where $\boldsymbol{y}_k$ and $\widehat{\boldsymbol{y}}_k$ are the ground-truth and predicted outputs for k^{th} task respectively. For unlabeled dataset, we leverage the contrastive (CL) loss function [24] to pull together the normalized representations of feature-view and visit-view of the same patient and to push apart these representations from representations of other patients.

$$
L_U = -\sum_{i=1}^{|D_k|} \sum_{\boldsymbol{z}_i \in \{\boldsymbol{z}_i^f, \boldsymbol{z}_i^v\}} \log \frac{\exp(\boldsymbol{z}_i^f \cdot \boldsymbol{z}_i^v)}{\sum_{\boldsymbol{z}_j \in A(\boldsymbol{z}_i)} \exp(\boldsymbol{z}_i \cdot \boldsymbol{z}_j)} \tag{17}
$$

where $A(\boldsymbol{z}_i) \equiv \boldsymbol{Z} \setminus \boldsymbol{z}_i$ in which $\boldsymbol{Z} = \{\boldsymbol{z}_i^f, \boldsymbol{z}_i^v\}_{i=1}^{|D_k|}$.

TABLE II: Cardiac complications in female breast cancer cohort and their corresponding ICD codes and numbers of positive instances.

complication	Description	ICD-10 Codes	#subjects
Atrial Fibrillation	An irregular, often rapid heart rate that commonly causes poor blood flow	I48	322
Coronary Artery Disease	Damage or disease in the heart's major blood vessels	I20-I25	769
Heart failure	A chronic condition in which the heart doesn't pump blood as well as it should	I11, I13 I42, I50	1124
Hypertension	A condition in which the force of the blood against the artery walls is too high	I10, I16	6787
Peripheral Arterial Disease	A circulatory condition in which narrowed blood vessels reduce blood flow to the limbs	I70	340
Stroke	Damage to the brain from interruption of its blood supply	I60-I69	592

IV. Experiments

In this section, we evaluate the performances of MuViTaNet on six real-world insurance claim datasets and compare its results with state-of-the-art clinical risk prediction models to demonstrate the effectiveness of our method. Besides achieving accurate prediction, we also show the robustness of MuViTaNet in terms of interpretability.

A. Datasets

Breast cancer cohort construction. We extract clinical records of female breast cancer patients from the MarketScan Commercial Claims and Encounter (CCAE) database provided by Truven Health[2] to construct cardiac complication risk profiling datasets. According to the previous work [19], the records from 2012 to 2017 of de-identified patients are selected based on the following criteria.

- Ages of the selected patients are from 18 to 65 at the initial diagnosis of breast cancer.
- The selected patients have at least six months of records and ten clinical visits before being diagnosed with breast cancer.
- There is no cardiac complication diagnosis until the initial diagnosis of breast cancer of the selected patients.

Cardiac complication datasets construction. After construing the breast cancer cohort, we create a distinct dataset for each cardiac complication onset prediction task. In our setting, we focus on profiling the risk of developing cardiac complications in a six-month window after the initial diagnosis of breast cancer, and the positive instances are defined as patients who have cardiac complications in this window. Following previous clinical research [3], [4], we identify six cardiac complications including atrial fibrillation (AF), coronary artery disease (CAD), heart failure (HF), hypertension, peripheral arterial disease (PAD), and stroke. Descriptions, ICD codes, and the corresponding numbers of positive instances of these complications are shown in Table II. The negative instances

[2]https://truvenhealth.com/markets/life-sciences/products/data-tools/marketscan-databases

TABLE III: Comparison of prediction performance measured by AU-ROC scores on six complication risk profiling tasks. We report the average AU-ROC scores and their corresponding standard deviation. (AF: Atrial Fibrillation, CAD: Coronary Artery Disease, HF: Heart Failure, PAD: Peripheral Arterial Disease).

	Method		AF	CAD	HF	Hypertension	PAD	Stroke	Average
Single-task	Classical	LR	0.6133 ± 0.0437	0.6402 ± 0.0165	0.6982 ± 0.0088	0.7901 ± 0.0088	0.5700 ± 0.0341	0.6150 ± 0.0128	0.6545 ± 0.0208
		RF	0.7159 ± 0.0434	0.7187 ± 0.0260	0.7863 ± 0.0147	0.8066 ± 0.0090	0.6880 ± 0.0525	0.7172 ± 0.0262	0.7388 ± 0.0286
	Recurrent-based	GRU	0.6701 ± 0.0425	0.7218 ± 0.0116	0.7805 ± 0.0033	0.8122 ± 0.0084	0.6884 ± 0.0368	0.7213 ± 0.0103	0.7324 ± 0.0188
		Bi-GRU	0.6620 ± 0.0533	0.7295 ± 0.0079	0.7845 ± 0.0058	0.8155 ± 0.0098	0.6967 ± 0.0172	0.7291 ± 0.0088	0.7362 ± 0.0171
	Time-aware	T-LSTM	0.6739 ± 0.0518	0.7052 ± 0.0133	0.7651 ± 0.0156	0.8024 ± 0.0118	0.6802 ± 0.0239	0.6994 ± 0.0203	0.7210 ± 0.0228
	Attention-based	Dipole	0.6804 ± 0.0661	0.7287 ± 0.0120	0.7791 ± 0.0026	0.8157 ± 0.0081	0.6839 ± 0.0320	0.7247 ± 0.0040	0.7354 ± 0.0209
		RETAIN	0.6493 ± 0.0465	0.6780 ± 0.0196	0.7360 ± 0.0139	0.8078 ± 0.0086	0.6731 ± 0.0224	0.6770 ± 0.0112	0.7035 ± 0.0126
		Transformer	0.6516 ± 0.0563	0.7021 ± 0.0155	0.7502 ± 0.0069	0.8107 ± 0.0076	0.6721 ± 0.0392	0.6981 ± 0.0135	0.7141 ± 0.0183
		LSAN	0.6069 ± 0.0556	0.6910 ± 0.0135	0.7567 ± 0.0180	0.8163 ± 0.0085	0.6464 ± 0.0464	0.6897 ± 0.0206	0.7012 ± 0.0271
Multi-task	Recurrent-based	GRU	0.7915 ± 0.0475	0.7759 ± 0.0144	0.8186 ± 0.0136	0.8143 ± 0.0096	0.7524 ± 0.0253	0.7458 ± 0.0222	0.7831 ± 0.0221
		Bi-GRU	0.7984 ± 0.0524	0.7824 ± 0.0121	0.8279 ± 0.0125	0.8189 ± 0.0100	0.7503 ± 0.0189	0.7462 ± 0.0237	0.7873 ± 0.0216
	Time-aware	T-LSTM	0.7944 ± 0.0466	0.7591 ± 0.0093	0.8134 ± 0.0124	0.8106 ± 0.0087	0.7382 ± 0.0285	0.7419 ± 0.0232	0.7763 ± 0.0214
	Attention-based	Dipole	0.7823 ± 0.0620	0.7814 ± 0.0213	0.8239 ± 0.0095	0.8210 ± 0.0092	0.7554 ± 0.0350	0.7611 ± 0.0194	0.7875 ± 0.0261
		RETAIN	0.7686 ± 0.0485	0.7554 ± 0.0083	0.8024 ± 0.0165	0.8029 ± 0.0066	0.7312 ± 0.0263	0.7376 ± 0.0254	0.7661 ± 0.0219
		Transformer	0.7697 ± 0.0649	0.7738 ± 0.0110	0.8049 ± 0.0164	0.8092 ± 0.0106	0.7484 ± 0.0423	0.7643 ± 0.0083	0.7784 ± 0.0256
		LSAN	0.7775 ± 0.0576	0.7788 ± 0.0225	0.8082 ± 0.0150	0.8226 ± 0.0061	0.7599 ± 0.0319	0.7533 ± 0.0147	0.7834 ± 0.0246
	Ours	MuViTaNet	$\mathbf{0.8120 \pm 0.0457}$	$\mathbf{0.8070 \pm 0.0147}$	$\mathbf{0.8408 \pm 0.0177}$	$\mathbf{0.8462 \pm 0.0089}$	$\mathbf{0.7986 \pm 0.0199}$	$\mathbf{0.7914 \pm 0.0174}$	$\mathbf{0.8160 \pm 0.0117}$

are randomly selected from the breast cancer cohort with a ratio of 3:1 compared to positive instances.

Unlabeled dataset construction. The negative patients that are not selected for complication datasets are used to construct a dataset for contrastive learning. MuViTaNet leverages this dataset as additional information to improve the prediction performances of complication onset prediction tasks.

Feature selection. We use the following information to profile cardiac complications for breast cancer patients.

- *Demographics including age and region information*. We cluster patients into three age groups (i.e., $18-44, 45-54, 55-65$) and five region groups.
- *Clinical codes including diagnosis, procedure, and medication codes*. For diagnosis codes, all ICD-9 codes are converted to ICD-10 codes. To alleviate data sparsity, we group all diagnosis and procedure codes based on their first three characters and remove codes that appear in less than 200 patients. For medication codes, we group them by their therapeutic classes. This preprocessing step results in 1188 features.

B. Experimental Setup

Baseline Models. To validate the performance of the proposed model for cardiac complication risk profiling task, we compare it with several state-of-the-art models. Based on their architectures, these models are categorized into four main groups including classical model, recurrent-based model, attention-based model, and time-aware model. The details of these models are presented as follows.

- **Logistic Regression (LR).** A classical model used in binary classification. To deal with insurance claim data, a patient record is converted to the count vector $\in \mathbb{Z}^{|\boldsymbol{C}|}$ whose i^{th} element is the frequency of i^{th} clinical code in that record, and is then fed into LR.
- **Random Forest (RF) [25].** A classical ensemble model whose prediction is the average computed from predictions of a number of decision tree classifiers. Inputs for RF are similar to LR.
- **Gated Recurrent Unit (GRU) [26].** A variant of recurrent neural network (RNN) that uses gating mechanism.
- **Bidirectional GRU (Bi-GRU) [20].** An improved version of GRU by employing an additional GRU model to learn the sequence data in reverse order.
- **Dipole [5].** An attention-based model that utilizes attention mechanism over the sequence generated by Bi-GRU to learn the dependencies between visits.
- **RETAIN [8].** An attention-based model that first employs a reverse RNN to process clinical records in reverse order to mimic physicians' decisions. Then two attention modules are used to identify significant visits and variables.
- **T-LSTM [6].** A time-aware model designed for handling irregularity visits in clinical records. The memory cell of LSTM is modified to capture time intervals between two consecutive visits.
- **Transformer [22].** A fully attention-based model that uses multi-head attention mechanisms to learn the dependencies among elements in sequential data.
- **LSAN [27].** An attention-based model that uses Transformer to capture global information and CNN to capture local information.
- **MTL Models**: We develop the MTL version for each of the aforementioned neural network-based models by employing task-specific attention and decoder over the output generated by these models.
- **MuViTaNet$^{\text{-visit-view}}$**: A variant of MuViTaNet by removing the visit-view encoder.
- **MuViTaNet$^{\text{-feature-view}}$**: A variant of MuViTaNet by removing the feature-view encoder.
- **MuViTaNet$^{\text{-task-specific}}$**: A variant of MuViTaNet by re-

TABLE IV: Top 10 most important clinical features (i.e., with the highest attention weights) for each cardiac complication as identified by MuViTaNet.

Atrial Fibrillation	Coronary Artery Disease	Heart Failure
Nonrheumatic mitral valve disorders (I34)	Other cardiac arrhythmias (I49)	Other cardiac arrhythmias (I49)
Other cardiac arrhythmias (I49)	Nonrheumatic mitral valve disorders (I34)	Varicose veins of lower extremities (I83)
Complications and ill-defined heart disease (I51)	Varicose veins of lower extremities (I83)	Diseases of capillaries (I78)
Paroxysmal tachycardia (I47)	Diseases of capillaries (I78)	Other disorders of veins (I87)
Diseases of capillaries (I78)	Type 2 diabetes mellitus (E11)	Embolism and thrombosis (I82)
Embolism and thrombosis (I82)	Other peripheral vascular diseases (I73)	Type 2 diabetes mellitus (E11)
Other conduction disorders (I45)	Embolism and thrombosis (I82)	Complications and ill-defined heart disease (I51)
Varicose veins of lower extremities (I83)	Hypotension (I95)	Nonrheumatic mitral valve disorders (I34)
Nonrheumatic aortic valve disorders (I35)	Other disorders of veins (I87)	Other peripheral vascular diseases (I73)
Other disorders of veins (I87)	Angina pectoris (I20)	Overweight and obesity (E66)
Hypertension	**Peripheral Arterial Disease**	**Stroke**
Other cardiac arrhythmias (I49)	Other cardiac arrhythmias (I49)	Other cardiac arrhythmias (I49)
Abnormal blood-pressure reading, without diagnosis (R03)	Varicose veins of lower extremities (I83)	Nonrheumatic mitral valve disorders (I34)
Type 2 diabetes mellitus (E11)	Diseases of capillaries (I78)	Varicose veins of lower extremities (I83)
Nonrheumatic mitral valve disorders (I34)	Nonrheumatic mitral valve disorders (I34)	Other peripheral vascular diseases (I73)
Varicose veins of lower extremities (I83)	Other disorders of veins (I87)	Embolism and thrombosis (I82)
Overweight and obesity (E66)	Nonspecific lymphadenitis (I88)	Type 2 diabetes mellitus (E11)
Diseases of capillaries (I78)	Other peripheral vascular diseases (I73)	Other disorders of veins (I87)
Other peripheral vascular diseases (I73)	Embolism and thrombosis (I82)	Hypotension (I95)
Other disorders of veins (I87)	Other noninfective disorders of lymphatic vessels (I89)	Pain in throat and chest (R07)
Pain in throat and chest (R07)	Type 2 diabetes mellitus (E11)	Complications and ill-defined heart disease (I51)

TABLE V: Average performances of MuViTaNet variants over 6 complication datasets (**F**: Feature-view, **V**: Visit-view, **L**: Labeled, **U**: Unlabeled).

Models	Multi-view		Multi-task		AU-ROC
	F	V	L	U	
MuViTaNet$^{\text{-task-specific}}$	✓	✓	✗	✗	0.7385 ± 0.0239
MuViTaNet$^{\text{-feature-view}}$	✗	✓	✓	✗	0.7906 ± 0.0286
MuViTaNet$^{\text{-visit-view}}$	✓	✗	✓	✗	0.7942 ± 0.0248
MuViTaNet$^{\text{-unlabeled}}$	✓	✓	✓	✗	0.8102 ± 0.0136
MuViTaNet	✓	✓	✓	✓	0.8160 ± 0.0117

moving the task-specific attention and decoder for single-task learning (STL) setting.

- **MuViTaNet$^{\text{-unlabeled}}$**: A variant of MuViTaNet trained with labeled datasets only.

Implementation Details. All neural network-based architectures are implemented by PyTorch[3]. For classical models including LR and RF, we use their Python implementations from Scikit-Learn [28]. We use ADAM algorithm [29] to optimize the prediction performances for neural network-based models. The batch size is set as 16 for labeled datasets and 256 for unlabeled dataset, and the initial learning rate is 0.0001.

Evaluation Metric. We conduct experiments under 5-fold cross-validation setting. 10% instances from the training set are used to construct the validation set, and the results on the testing set are determined based on the best results on the validation set. The area under the receiver operating characteristic (AU-ROC) is used to measure the performances of prediction models for cardiac complication risk profiling.

C. Results

We conduct experiments to answer the following questions.

- **Q1.** How accurate is MuViTaNet for cardiac complication risk profiling task comparing to previous works?
- **Q2.** How each component of MuViTaNet contributes to its prediction performance?
- **Q3.** How to effectively interpret the predictions made by MuViTaNet?

Cardiac complication risk profiling. As shown in Tables III, MuViTaNet achieves the best performances compared to other baselines for cardiac complication risk profiling task measured by AU-ROC score. Generally, it achieves an average (i.e., over six datasets) AU-ROC score of 0.8102, which is 11% better than the best previous method. Looking into each complication dataset, we also observe that MuViTaNet consistently outperforms other methods in terms of AU-ROC score. Such improvements indicate the advantage of MuViTaNet by using (1) multi-view encoder to extract comprehensive information and (2) MTL scheme to leverage information from both related labeled and unlabeled datasets to improve its prediction performance.

For baseline methods, we can observe that formulating complication risk profiling as MTL significantly improves the prediction performances of these methods. The improvements are more noteworthy for small datasets, including AF (31%), CAD (19%), PAD (22%), and stroke (13%). These results demonstrate the importance of leveraging task-related information for predicting the onset of complications. We also see that GRU-based models achieve slightly improved performances compared to other neural network models. For STL setting, the averaged prediction performances of deep learning models are on par with RF and are much better than LR. To investigate more, we zoom into the prediction performance for each dataset and observe that RF outperforms deep learning models for AF, CAD, PAD, and stroke datasets whose sizes are relatively small compared to HF and hypertension datasets. This result is reasonable because deep learning methods generally require large training data to achieve good prediction performance.

Ablation study. To investigate the contribution of each component in MuViTaNet, we conduct an ablation study by comparing MuViTaNet with its simpler variants including MuViTaNet$^{\text{-visit-view}}$, MuViTaNet$^{\text{-feature-view}}$,

[3]https://pytorch.org/

TABLE VI: Top 5 most important clinical visits and features (i.e., with the highest attention weights) for the 2 patients illustrated in Figure 4.

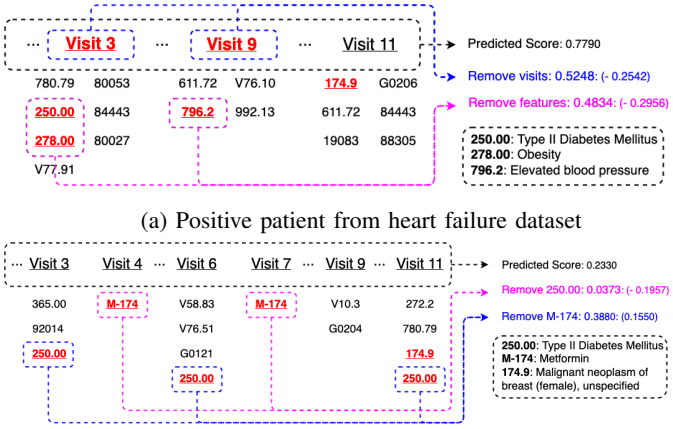
Positive patient from heart failure dataset					
Visits	Visit 9 (0.11)	Visit 3 (0.11)	Visit 11 (0.10)	Visit 8 (0.09)	Visit 6 (0.09)
Features	796.2 (0.26)	250.00 (0.25)	278.00 (0.12)	882.0 (0.05)	19083 (0.04)
Negative patient from hypertension dataset					
Visits	Visit 9 (0.11)	Visit 11 (0.11)	Visit 7 (0.10)	Visit 4 (0.10)	Visit 3 (0.09)
Features	M-174 (0.56)	250.00 (0.22)	S0612 (0.13)	J3010 (0.02)	82043 (0.02)

MuViTaNet$^{\text{-task-specific}}$, and MuViTaNet$^{\text{-unlabeled}}$ on the six aforementioned datasets. The AU-ROC scores of these models are shown in Table V. We can observe that encoding clinical data solely by a single-view encoder is not as good as a multi-view encoder. AU-ROC score of MuViTaNet decreases to 0.7906 (resp. 0.7942) when only using visit-view (resp. feature-view) encoder. This result demonstrates the necessity of aggregating information from multiple views. The performance of MuViTaNet also drops significantly when we remove the task-specific attention mechanism and decoder, which further confirms the importance of formulating complication risk profiling task as MTL with both labeled and unlabeled datasets.

Model interpretability. The deployment of data-driven systems to healthcare applicants in real-world requires not only models with good prediction performance but also efficient mechanisms to interpret the automated decision to clinicians. By leveraging the multi-view multi-task architecture, our proposed model can interpret the prediction for each complication in multiple perspectives, thereby helping clinicians understand which clinical entities contribute most to the prediction.

To characterize cardiac complications, we find the most important features for each of these cardiac complications by averaging the feature-view attention weights over all positive patients for clinical features in each complication dataset. Due to the varied number of features across patients, we rescale attention weights by multiplying them with the number of features appeared in the corresponding records before averaging. Then top-10 clinical features for 6 cardiac complications are shown in Table IV. We observe that these complications share many common features such as **I34** (Nonrheumatic mitral valve disorders), **I49** (Other cardiac arrhythmias), etc. This result is reasonable because all of these complications belong to cardiovascular disease class. Moreover, many important features determined by our model are known to be clinically associated with the corresponding complications. For example, patients with type II diabetes are two to four times more likely to develop heart diseases than someone without diabetes [30]. Obesity is another major known risk factor for heart failure and hypertension patients [31], [32]. Angina pectoris is the type of chest pain caused by reduced blood flow to the heart and is considered as a symptom of coronary artery disease [33].

Case study. To further investigate the interpretability of MuViTaNet, we look at two case studies to visualize the learned attention weights for finding risk factors of each complication. The case studies include a positive patient from heart failure dataset and a negative patient from hypertension dataset. Their clinical records are illustrated in Figure 4. The most important visits and features determined by their associated attention weights from visit-view and feature-view task-specific attention components are shown in Table VI. For the positive patient (Figure 4a), the predicted probability for heart failure onset is 0.7790. As shown in Table VI, the visit-view attention focuses more on visits 3 and 9, which include clinical codes **250.00** (Type II diabetes mellitus) and **278.00** (Obesity) and these codes are also determined as the most important features by the feature-view attention. This result is also consistent with clinical research in which type II diabetes mellitus and obesity have been shown as the common risk factors for heart failure disease [30], [32], thereby demonstrating the effectiveness of MuViTaNet in capturing the correlation between risk factors and corresponding diseases. To further investigate the robustness of our model, we remove important visits and features indicating heart failure's risk factors from the patient record and predict the probability of heart failure onset based on the modified records for capturing the changes in model output. Figure 4a shows that the predicted score decreases to 0.5284 and 0.4834 when removing visits (3 and 9) and codes (**250.00**, **278.00**, and **796.2**) respectively. Thus, MuViTaNet is capable to focus on clinical-related visits and features when predicting onset of complications.



(a) Positive patient from heart failure dataset

(b) Negative patient from hypertension dataset

Fig. 4: Visualization of 2 patient records (i.e., positive patient from heart failure dataset and negative patient from hypertension dataset) from breast cancer cohort. We only show important visits in clinical records due to limited space.

Figure 4b shows a clinical record of the negative patient who has type II diabetes mellitus but is also treated by **M-174** (Metformin). Tables VI indicates that MuViTaNet pays more attention on **M-174** and **250.00** when predicting onset of hypertension. To verify whether our model can capture the relationship between disease and treatment, we remove these codes from the patient record as we did for the positive patient. Figure 4b shows that the predicted probability increases from 0.2330 to 0.3380 when removing Metformin (diabetes medication) and decreases to 0.0373 when removing code **250.00** (diabetes). This result indicates that MuViTaNet considers

the impact of both disease and treatment on complication development when making predictions.

V. Conclusions

Complication risk profiling is a crucial problem in healthcare prediction domain. In this paper, we propose a novel multi-view multi-task network (MuViTaNet) that leverages clinical data to profile multiple complications for patients. To tackle the issues of existing methods, MuViTaNet considers the record as the sequence of clinical visits and the set of clinical features, and then employs the multi-view encoder to effectively extract meaningful information from both feature-view and visit-view of the patient record. Due to the relatedness among different complications, we organize MuViTaNet as the MTL architecture in which the shared representation learned from the multi-view encoder is put into multiple task-specific attention components to learn task-specific representations for patients in both labeled and unlabeled datasets. Finally, the predicted probability for each complication onset is generated from the task-specific representation by the corresponding decoder. We evaluate the prediction performance of MuViTaNet on the insurance claim database which consists of 6 cardiac complication datasets for breast cancer survivors. The experimental results demonstrate that our proposed model outperforms other state-of-the-art models for the complication risk profiling task. More importantly, MuViTaNet provides an efficient mechanism to interpret their prediction from multiple perspectives, thereby helping clinicians to make better decisions in real-world scenarios.

Acknowledgment

This work was funded in part by the National Science Foundation under award number CBET-2037398.

References

[1] C. Schairer, P. J. Mink, L. Carroll, and S. S. Devesa, "Probabilities of death from breast cancer and other causes among female breast cancer patients," *Journal of the National Cancer Institute*, vol. 96, no. 17, 2004.

[2] J. L. Patnaik, T. Byers, C. DiGuiseppi, D. Dabelea, and T. D. Denberg, "Cardiovascular disease competes with breast cancer as the leading cause of death for older females diagnosed with breast cancer: a retrospective cohort study," *Breast Cancer Research*, vol. 13, no. 3, 2011.

[3] H. Abdel-Qadir, P. Thavendiranathan, K. Fung, E. Amir, P. C. Austin, G. S. Anderson, and D. S. Lee, "Association of early-stage breast cancer and subsequent chemotherapy with risk of atrial fibrillation," *JAMA network open*, vol. 2, no. 9, 2019.

[4] H. Strongman, S. Gadd, A. Matthews, K. E. Mansfield, S. Stanway, A. R. Lyon, I. dos Santos-Silva, L. Smeeth, and K. Bhaskaran, "Medium and long-term risks of specific cardiovascular diseases in survivors of 20 adult cancers: a population-based cohort study using multiple linked uk electronic health records databases," *The Lancet*, vol. 394, no. 10203, 2019.

[5] 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 *KDD'17*, 2017.

[6] I. M. Baytas, C. Xiao, X. Zhang, F. Wang, A. K. Jain, and J. Zhou, "Patient subtyping via time-aware lstm networks," in *KDD'17*, 2017.

[7] J. Gao, C. Xiao, Y. Wang, W. Tang, L. M. Glass, and J. Sun, "Stagenet: Stage-aware neural networks for health risk prediction," in *WWW'20*, 2020.

[8] E. Choi, M. T. Bahadori, J. A. Kulas, A. Schuetz, W. F. Stewart, and J. Sun, "Retain: An interpretable predictive model for healthcare using reverse time attention mechanism," in *NIPS'16*, 2016.

[9] H. Song, D. Rajan, J. Thiagarajan, and A. Spanias, "Attend and diagnose: Clinical time series analysis using attention models," in *AAAI'18*, vol. 32, no. 1, 2018.

[10] T. Bai, S. Zhang, B. L. Egleston, and S. Vucetic, "Interpretable representation learning for healthcare via capturing disease progression through time," in *KDD'18*, 2018.

[11] B. C. Kwon, M.-J. Choi, J. T. Kim, E. Choi, Y. B. Kim, S. Kwon, J. Sun, and J. Choo, "Retainvis: Visual analytics with interpretable and interactive recurrent neural networks on electronic medical records," *IEEE transactions on visualization and computer graphics*, vol. 25, no. 1, 2018.

[12] L. Ma, C. Zhang, Y. Wang, W. Ruan, J. Wang, W. Tang, X. Ma, X. Gao, and J. Gao, "Concare: Personalized clinical feature embedding via capturing the healthcare context," in *AAAI'20*, vol. 34, no. 01, 2020.

[13] J. Zhou, L. Yuan, J. Liu, and J. Ye, "A multi-task learning formulation for predicting disease progression," in *KDD'11*, 2011.

[14] B. Liu, Y. Li, Z. Sun, S. Ghosh, and K. Ng, "Early prediction of diabetes complications from electronic health records: A multi-task survival analysis approach," in *AAAI'18*, vol. 32, no. 1, 2018.

[15] J. Wiens, J. Guttag, and E. Horvitz, "Patient risk stratification with time-varying parameters: a multitask learning approach," *The Journal of Machine Learning Research*, vol. 17, no. 1, 2016.

[16] N. Nori, H. Kashima, K. Yamashita, H. Ikai, and Y. Imanaka, "Simultaneous modeling of multiple diseases for mortality prediction in acute hospital care," in *KDD'15*, 2015.

[17] N. Razavian, J. Marcus, and D. Sontag, "Multi-task prediction of disease onsets from longitudinal laboratory tests," in *MLHC'16*. PMLR, 2016.

[18] Z. C. Lipton, D. C. Kale, C. Elkan, and R. Wetzel, "Learning to diagnose with LSTM recurrent neural networks," in *ICLR'16*, 2016.

[19] B. Liu, Y. Li, S. Ghosh, Z. Sun, K. Ng, and J. Hu, "Complication risk profiling in diabetes care: A bayesian multi-task and feature relationship learning approach," *IEEE Transactions on Knowledge and Data Engineering*, vol. 32, no. 7, 2019.

[20] B. Ljubic, A. A. Hai, M. Stanojevic, W. Diaz, D. Polimac, M. Pavlovski, and Z. Obradovic, "Predicting complications of diabetes mellitus using advanced machine learning algorithms," *Journal of the American Medical Informatics Association*, vol. 27, no. 9, 2020.

[21] A. Guo, K. W. Zhang, K. Reynolds, and R. E. Foraker, "Coronary heart disease and mortality following a breast cancer diagnosis," *BMC medical informatics and decision making*, vol. 20, 2020.

[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *NIPS'17*, 2017.

[23] D. Dong, H. Wu, W. He, D. Yu, and H. Wang, "Multi-task learning for multiple language translation," in *ACL'15*, 2015.

[24] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *ICML'20*. PMLR, 2020.

[25] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, 2001.

[26] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," in *EMNLP'14*, 2014.

[27] M. Ye, J. Luo, C. Xiao, and F. Ma, "Lsan: Modeling long-term dependencies and short-term correlations with hierarchical attention for risk prediction," in *CIKM'20*, 2020.

[28] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, 2011.

[29] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR'15*, 2015.

[30] H. C. Kenny and E. D. Abel, "Heart failure in type 2 diabetes mellitus: impact of glucose-lowering agents, heart failure therapies, and novel therapeutic strategies," *Circulation research*, vol. 124, no. 1, 2019.

[31] N. Mikhail, M. S. Golub, and M. L. Tuck, "Obesity and hypertension," *Progress in cardiovascular diseases*, vol. 42, no. 1, 1999.

[32] I. A. Ebong, D. C. Goff Jr, C. J. Rodriguez, H. Chen, and A. G. Bertoni, "Mechanisms of heart failure in obesity," *Obesity research & clinical practice*, vol. 8, no. 6, 2014.

[33] M. Mosseri, R. Yarom, M. Gotsman, and Y. Hasin, "Histologic evidence for small-vessel coronary artery disease in patients with angina pectoris and patent large coronary arteries." *Circulation*, vol. 74, no. 5, 1986.