Title: Task-Oriented Predictive (Top)-BERT: Novel Approach for Predicting Diabetic Complications Using a Single-Center EHR Data

Author List: Humayera Islam[1], MS, Gillian Bartlett[1,2,4], PhD, Robert Pierce[4], MD, Praveen Rao[1,3], PhD, Lemuel R. Waitman[1,2,4], PhD, Xing Song[1,2], PhD

Institutions: [1]Institute for Data Science and Informatics, [2]Department of Biomedical Informatics, Biostatistics, and Medical Epidemiology, [3]Department of Electrical Engineering and Computer Science, [4]University of Missouri School of Medicine, Columbia, USA

Abstract

In this study, we assess the capacity of the BERT (Bidirectional Encoder Representations from Transformers) framework to predict a 12-month risk for major diabetic complications—retinopathy, nephropathy, neuropathy, and major adverse cardiovascular events (MACE) using a single-center EHR dataset. We introduce a task-oriented predictive (Top)-BERT architecture, which is a unique end-to-end training and evaluation framework utilizing sequential input structure, embedding layer, and encoder stacks inherent to BERT. This enhanced architecture trains and evaluates the model across multiple learning tasks simultaneously, enhancing the model's ability to learn from a limited amount of data. Our findings demonstrate that this approach can outperform both traditional pretraining-finetuning BERT models and conventional machine learning methods, offering a promising tool for early identification of patients at risk of diabetes-related complications. We also investigate how different temporal embedding strategies affect the model's predictive capabilities, with simpler designs yielding better performance. The use of Integrated Gradients (IG) augments the explainability of our predictive models, yielding feature attributions that substantiate the clinical significance of this study. Finally, this study also highlights the essential role of proactive symptom assessment and the management of comorbid conditions in preventing the advancement of complications in patients with diabetes.

Introduction

Micro and macro-vascular complications induced by diabetes can have substantial impact on diabetes management and patient care[1,2]. Early prediction of these complications allows for the identification of high-risk patients and active implementation of preventive measures[3–7]. With this motivation, researchers have developed models predicting diabetes-related complications, primarily emphasizing cardiovascular outcomes and, to a lesser extent, kidney and eye complications[3,8]. However, most of the prior research was focused on predicting risk scores using a limited number of risk factors, often curated from previous literature [8–14]. Despite many machine learning (ML) and deep learning (DL) models that emerged in recent research, classical ML models dominated these studies– mostly limited to performance comparisons, with only a minority delving into exploring novel risk factors and discovering new knowledge[15,16].

Digital patient data from electronic health records (EHR) systems play a crucial role in developing clinical risk prediction models, thereby guiding the development of robust, evidence-based medical interventions[13,17,18]. Structured EHR systems systematically documents the timeline of patient encounters, encompassing elements such as demographics, vital signs, diagnoses, prescribed medications, lab test results, and medical procedures. Hence, feature vectors derived from EHR data can enable the use of traditional ML and DL techniques[19,20]. However, the intricate and abundant information contained within EHR data is often condensed to create summary features for predictive models. This process can diminish the temporal and contextual richness of the data. This simplification frequently neglects the complex nature of EHR data, such as sparsity, heterogeneity, and irregular patterns of patient visits, leading to model overfitting and lack of model generalizability[21].

The resemblance between EHR sequences (time series sequences from different data modalities) and natural language (word sequences) has led to the adoption of advanced NLP (natural language processing) techniques for EHR data. Convolutional neural networks (CNN)[22] and recurrent neural networks (RNN)[23] with embedding layers have enhanced the capture of sequential data but have fallen short in recognizing long-term dependencies, hindered by problems like gradient instability. Subsequent techniques, such as Long Short-Term Memory (LSTM) models, have been designed to forecast clinical events, yet they have been hampered by slow training processes and persistent data complexities[24,25]. A transformative breakthrough that revolutionized the learning of contextual and temporal information in language models is the Transformers (2017)[26] architecture. The pioneering studies[21,27,28] in applying Transformers to structured EHR data, notably BERT (Bidirectional Encoder Representations from Transformers, 2019[29]), have shown their effectiveness in capturing the complex temporal patterns and navigating the intricacies of EHR for clinical predictions. For instance, BERT's advanced embedding framework captures the nuanced semantics and context of patient timelines, addressing EHR data sparsity and varying time intervals between encounters. Unlike traditional ML models, BERT transforms these sequences into dense embeddings, preventing learning from sparse matrices of numerous zeros. Moreover, its self-attention and feed-forward mechanisms efficiently learn long-term dependencies and uncover complex event relationships, improving model transparency.

Earlier studies adopted a two-stage process where pretraining on extensive multi-site EHR databases was followed by finetuning on a smaller, specific cohort for clinical predictions. For example, Med-BERT (2021)[27] was pretrained on a multi-site dataset encompassing 28 million patient records for one week before finetuning on three smaller, distinct datasets. However, patient privacy laws and proprietary data rights present significant obstacles to data sharing[30,31], thus impeding the distribution of EHR-based pretrained models, deviating from a common practice in the NLP field. This leads us to the question: Could BERT's unique ability to manage EHR-specific complexities still provide valuable and interpretable predictions when applied to data from a single medical center?

Therefore, in this study, we investigated BERT's potential in predicting a 12-month risk of developing significant complications, including retinopathy (RET), chronic kidney disease (CKD), neuropathy (NEUR), and major adverse cardiovascular events (MACE) in diabetes patients from a single-center EHR data. Shifting from the conventional pretraining-finetuning paradigm, we introduced an end-to-end training and evaluation method called task-oriented-predictive (Top)-BERT. This innovative approach concurrently optimizes the model for multiple specific prediction tasks—enhancing its effectiveness particularly in settings constrained by limited data.

In Top-BERT, we utilized the sequential input structure, embedding layer, and encoder stacks inherent to BERT to train and evaluate three tasks simultaneously: the conventional Masked Language Model (MLM), a binary classification for prolonged hospital stay (1 if the length of stay >7, else 0), and a multilabel sequence classification for the four complications mentioned above. We aggregated the loss of the three tasks, which was backpropagated throughout the entire network, leading to improved learning of our model in a limited cohort sample size. We evaluated our Top-BERT model against conventional pretraining-finetuning experiments with sequential input. We also compared its performance with traditional ML techniques, such as XGBoost, using a one-hot encoded representation of the features. Our Top-BERT model demonstrated a more effective ability to distinguish between classes and maintained robustness in managing the class imbalance for multilabel outputs, outperforming both the traditional approaches of pretraining-finetuning and XGBoost models.

We also investigated into the embedding structure of BERT, which typically uses positional and segment embeddings, to discern the sequential order of patient histories. Prior studies implemented unique embeddings, such as Med-BERT's incorporation of visit numbers[27] and BEHRT's addition of age to

positional and segment embeddings[21]. Our study evaluated the impact of integrating temporal factors as embeddings—such as age, visit sequence, and inter-visit intervals—to predict diabetes complications. We utilized AUROC (area under the receiver operating characteristics curve) and Shannon's entropy [32] for performance comparison, aiming to determine the most informative temporal representation in patient histories for our single-center dataset.

To the best of our knowledge, our study represents a novel effort to apply a modified BERT architecture for predicting four significant micro and macro-vascular complications in diabetes patients within a single research framework while navigating the complexities inherent in EHR data. Furthermore, in our study, we systematically evaluate feature importance across patient visits, identify highly contributing features for each complication, and examine age-specific feature influence variations using Integrated Gradients (IG), a gradient-based feature attribution method optimized for deep learning[33]. Unlike attention mechanisms in models like BERT, which offer partial insights, IG provides a comprehensive analysis by tracing the gradient flow from a predefined baseline to the input. This approach ensures adherence to the axioms of sensitivity and implementation invariance, which are not satisfied by other backpropagation methods such as layer-wise relevance propagation (LRP)[34] or Deconvolutional networks (DeConvNets)[35]. This methodological adaptation significantly enhances the explainability of our study's findings within the clinical setting, offering nuanced insights into the factors driving the model's predictions.

The remainder of this paper is structured as follows. The method section outlines our approach to model derivation and development, data preparation, followed by the detailed experimental setup for this study. The result section presents the comprehensive analysis of the study cohort, model comparison, and analysis of the model explanations. Finally, the discussion section elucidates the methodological advances and clinical implications of our findings and provides a concluding summary at the end.

Methods

Model Derivation & Development

Conventional BERT framework. The architecture of BERT[29], initially designed for language representation, is built upon a multi-layer bidirectional Transformer encoder based on Vaswani (2017)[36]. The fundamental elements of BERT encompass (i) input/output representation, (ii) the configuration of embedding layers, and (iii) the architecture of the Transformer encoder layers. In essence, BERT necessitates input sequencing, which involves tokenization using a designated vocabulary and incorporating special tokens like [CLS] at the beginning of the sequence and [SEP] to denote sequence separation. After tokenization, the input token traverses through the embedding layers, each capturing distinct contextual facets of the sequence. This process yields a summed embedding comprising token, segment, and positional embeddings. The embedded input then passes through the Transformer Encoder stack, processing every token simultaneously.

BERT training entails two steps: (i) pretraining tasks and (ii) finetuning tasks. Pretraining involves two self-supervised tasks. The Masked Language Model (MLM) randomly masks a fraction of input tokens and subsequently predicts these masked tokens through a training head atop the encoder stack. Concurrently, Next Sentence Prediction (NSP), a binary classification task, further trains BERT to comprehend inter-sequence relationships. During pretraining, contextualized embeddings are generated for each input token. In the finetuning phase, the pretraining weights are loaded to train the finetuning cohort for specific downstream prediction tasks like classification with an additional prediction head (classification layer) integrated over the encoder stack.

Motivation for the input representation for EHR data in BERT. The patient timeline in an EHR is a series of visits, each documented with various health-related elements such as diagnoses, medications, and

lab results. Reflecting this, Li et al. (2020)[21] designed BEHRT, inspired by BERT, using patient diagnoses information per visit to predict future diagnoses, denoting the EHR timeline of each patient p with n_p number of visits as, $V_p = \{\boldsymbol{v}_,^1\ \boldsymbol{v}_,^2\ \boldsymbol{v}_,^3 \dots, \boldsymbol{v}^{n_p}\}$, where $\boldsymbol{v}^j = \{v_1^j, v_2^j, \dots, v_m^j\}$ contains ordered clinical entities in the jth visit. Similar to the BERT model, they introduced the start of medical history (i.e., [CLS]) and the space between visits (i.e., [SEP]), which results in a new sequence, $V_p = \{[CLS], \boldsymbol{v}_,^1\ [SEP], \boldsymbol{v}_,^2\ [SEP], \boldsymbol{v}_,^3 \dots, \boldsymbol{v}^{n_p}, [SEP]\}$. However, Med-BERT(2021)[27] did not use the specific tokens [CLS] and [SEP] at the input layer due to differences in EHR and text input formats. In BERT, [SEP] serves as a separator between two adjacent sentences for the next sentence prediction task, and as reasoned for Med-BERT, visit embeddings effectively separate each visit in EHR, making the addition of [SEP] redundant. Similarly, Rao-BEHRT (2022)[28] modified the BEHRT representation of EHR sequences and dropped the [CLS] token. In all these studies, the visit sequence for each patient, which consists of diagnoses/medications, is converted to sequence to represent the temporal structure of EHR.

Moreover, adaptations of the embedding layer for EHR temporality representation varied in the previous studies. The embedding layer in BEHRT[21] incorporates four types of embeddings: disease, position, age, and visit segment. Positional encodings enable the network to capture positional interactions among diseases using a pre-determined encoding addressing the imbalanced distribution of sequence length in EHR. Age serves as a risk factor for diseases and provides chronological information, while visit segment indicates the separation between visits and differentiates adjacent visits of a patient. Similarly, Med-BERT (2021)[27] utilized diagnosis code embeddings, visit embeddings, and serialization embeddings to capture clinical code representations, distinguish visits, and capture code order. Additionally, Rao-BEHRT (2022)[28] used encounter (disease/medication), age, and calendar year. Moreover, Med-BERT was trained on structured diagnosis data using ICD codes, unlike BEHRT and Rao-BEHRT (2022)[28], which used Caliber codes (developed by a college in London).

Our proposed Top-BERT architecture. Our designed Top-BERT, leveraging the foundational components of BERT, serves as a versatile end-to-end training and evaluation architecture that can be tailored directly for a wide range of clinical predictive tasks. Top-BERT utilizes input representation, embedding layer, and encoder stacks like traditional BERT. The input sequence for Top-BERT represents the EHR sequence for each patient p with n_p number of visits as: $V_p = \{[CLS], \boldsymbol{v}^1, \boldsymbol{v}^2, \boldsymbol{v}^3 \dots, \boldsymbol{v}^{n_p}\}$, where $\boldsymbol{v}^j = \{v_1^j, v_2^j, \dots, v_m^j\}$ contains the ordered clinical entities in the jth visit . We added the [CLS] token at the start of every sequence, which is essential for training BERT for classification tasks. During training, the final hidden state of the [CLS] token becomes the summary representation of the sequence. Due to BERT's self-attention layers, the [CLS] token integrates context from the full sequence, resulting in a complete summary by the last layer, making it suitable for sequence-level classification tasks. In Top-BERT's embedding architecture, we examined different temporal factors such as age, number of visits, time between visits, and the conventional positional and segment embeddings to best represent patient timelines in EHRs. The specifics of these experiments will be elaborated in our experimental design section.

Figure 1 shows the modified BERT architecture used for training Top-BERT using the conventional terminology of the model. In the standard BERT pretraining[29], the input embeddings for ith token in jth visit, v_i^j represented as $\boldsymbol{e}_i \in \mathbb{R}^{hidden\_size}$, passing through the $BERT$ model, get transformed into context vectors and then into hidden states by the BertAttention layer. These are further processed by BertIntermediate's dense layer and normalized by BertOutput. The BertEncoder stack, comprising multiple BertLayers, yields encoded outputs for all encoder layers (*all_encoded_layers*). The BertModel produces two primary outputs: *all_encoded_layers* and *pooled_output*, the latter derived from the BertPooler function applied to the hidden state of the initial [CLS] token. The *pooled_output* is essential for training the model for any classification tasks by adding an appropriate dense layer to output logits.

In the MLM task, for a masked token v_{mask}^{j}, the model output before the activation layer can be represented as:

$$\boldsymbol{h}_{mask} = BERT(\boldsymbol{e}_{mask}) \text{ and}$$
$$\boldsymbol{o}_{m} = softmax(\boldsymbol{W}\boldsymbol{h}_{mask} + \boldsymbol{b}),$$

where $\boldsymbol{e}_{mask}$ is the embedding of the masked input, $\boldsymbol{h}_{mask}$ is the output of $BERT$ model for the masked token, $\boldsymbol{W}$ is the weight matrix of the output layer, $\boldsymbol{b}$ is the bias term, and $\boldsymbol{o}_{m}$ is the predicted probabilities.

Similarly, for the NSP task, for a token v_{i}^{j}, the model output is represented as:

$$\boldsymbol{h}_{nsp} = BERT(\boldsymbol{e}_{i}) \text{ and}$$
$$\boldsymbol{o}_{nsp} = \sigma(\boldsymbol{W}\boldsymbol{h}_{nsp} + \boldsymbol{b}),$$

where $\boldsymbol{e}_{i}$ is the embedding of the input, $\boldsymbol{h}_{nsp}$ is the hidden state of [CLS] token (*pooled_output)*, $\boldsymbol{W}$ and $\boldsymbol{b}$ are the weights and biases for NSP, and $\boldsymbol{o}_{nsp}$ is the predicted probabilities. $\sigma(\cdot)$ is the sigmoid activation function.

As a novelty of our approach, we have augmented the BERT architecture with a sequence classification head, merging the pretraining and finetuning steps into one end-to-end training and evaluation process. While the original pretraining heads focus on BERT's standard tasks, the new head utilizes the *pooled_output* to yield precise logits for sequence classification (dense layer to output logits). For this sequence classification task, we define the output as:

$$\boldsymbol{h}_{seq} = BERT(e_{i}) \text{ and}$$
$$\boldsymbol{o}_{seq} = \sigma(\boldsymbol{W}h_{seq} + \boldsymbol{b}),$$

where $\boldsymbol{h}_{seq}$ is the hidden state of [CLS] token (*pooled_output)* and o_{seq} is the predicted probabilities. Thus, Top-BERT can be represented as:

$$topBERT\left(v_{i}^{j}\right) = f(\boldsymbol{h}_{mask}, \boldsymbol{h}_{nsp}, \boldsymbol{h}_{seq}),$$

where $f(\cdot)$ represents the multitask model with self-supervised ($\boldsymbol{h}_{mask}$) and semi-supervised ($\boldsymbol{h}_{nsp}$), and supervised ($\boldsymbol{h}_{seq}$) tasks.

This enhancement aligns with BERT's original design and extends its utility by introducing multitasking capabilities within the pretraining phase itself. Each task's logits undergo their respective loss functions, and the cumulative loss from the three tasks is backpropagated, updating the network's weights. The combined loss can be represented as:

$$\mathcal{L} = \sum_{t} \ell_{t},$$

where $t \in \{MASK,\ NSP, SEQ\}$ and each ℓ_{t} can be represented as a summation of each task-oriented cross-entropy loss:

$$\ell_{t} = -\sum_{c=1}^{C} \boldsymbol{y}_{o,c} \log(\boldsymbol{o}_{t,c}),$$

where C is the number of classes, y_{o} is the one-hot encoded true label, and $o_{t,c}$ is the predicted probability for class c.

This results in training and evaluating the BERT model for specific prediction tasks, refining its capability by simultaneously optimizing it for multiple objectives. This multitask optimization could be pivotal in domains with limited data or where task-specific nuances are critical, offering a more nuanced and direct path to task-specific model refinement. We will discuss the implementation of Top-BERT for diabetes-related complication prediction tasks in the following subsections.

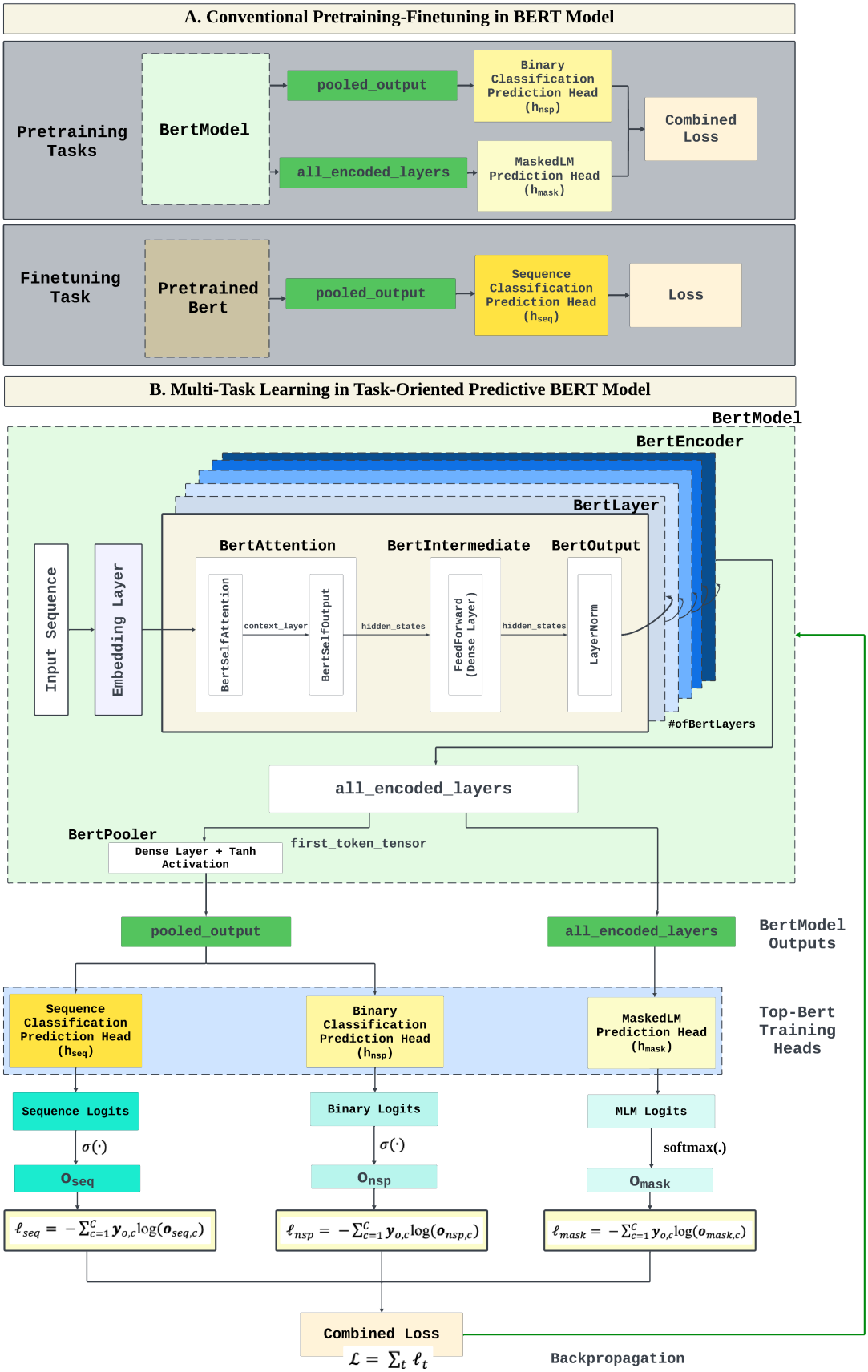


Figure 1A. This figure shows the conventional pretraining-finetuning steps in BERT framework. During pretraining, BERT employs two self-supervised tasks: Masked Language Model (MLM) and Next Sentence Prediction (NSP), which discerns relationships between sentence pairs. For finetuning, these pretrained weights initialize a model tailored to specific tasks. Figure 1B. An infographic showing the task-oriented predictive BERT framework using the conventional terminology in BERT models. In BERT's architecture, embedded inputs pass through BertAttention layers to generate context-sensitive hidden states, further refined by a series of BertLayers to produce a stack of encoded outputs. The model outputs both these encoded layers and a pooled_output–the latter obtained from the hidden state of [CLS] token. We enhanced the BERT architecture and blended pretraining and finetuning into a unified process by integrating a sequence classification head. Utilizing pooled_output from the BertPooler, our adapted model conducts end-to-end training for sequence classification tasks. This method leverages multitasking during pretraining, allowing for simultaneous loss optimization across tasks and enhancing the model's predictive performance.

Data Preparation

Data Source. Our research utilized the EHR database from the University of Missouri (MU) Hospital, which contains over 1.25 million patient records reflecting a wide geographic and demographic diversity range. MU follows the PCORnet Common Data Model (CDM) for structuring and representing their EHR data. This model employs standardized vocabularies such as Systematized Nomenclature of Medicine-Clinical Terminology (SNOMED CT), Current Procedural Terminology (CPT), and the ICD (International Classification of Diseases) versions 9 and 10 for consistent data mapping. The deidentified version of the MU CDM, updated in October 2022 with altered dates and pseudo-identifiers, served as the foundation for our study. Database queries were executed using the Snowflake computing platform. The MU Institutional Review Board (IRB) approved this research.

Cohort identification. We utilized the framework from Furmanchuk (2021)[37] to identify the diabetes mellitus (DM) cohort from our MU CDM based on the SUrveillance, PREvention, and ManagEment of Diabetes Mellitus (SUPREME-DM) algorithm. the SUPREME-DM DataLink is one example of a distributed registry developed for studying *Any*-DM (mixed Type 1 DM and Type 2 DM codes) using a standardized data extraction approach based on diagnosis, labs, and medications[38–40]. Although SUPREME-DM has not focused on distinguishing adults with Type I DM vs. Type II DM, this algorithm has been shown to have the potential of extracting the most representative EHR-based DM cohort[40]. We defined the MU study denominator as any patient between ages 18 to 89 years at the visit with at least two distinct encounter days. The encounter types included ambulatory visit (AV), emergency department (ED), emergency department admit to inpatient hospital stay (EI), inpatient hospital stays (IP), non-acute institutional stay (IS), and telehealth (TH) between 01/01/2010 and 01/31/2023.

We implemented the definition of SUPREME-DM on the MU denominator using the following steps (as detailed in Figure 2A): Exclusion based on periods of pregnancy. We first excluded pregnancy-related encounters using relevant ICD and CPT codes, then masked encounters within a year of each identified pregnancy. Diagnosis codes. We identified diabetic patients as those having two visits with diabetes-related ICD codes on separate days within two years, noting the date of the initial visit. Lab codes. Using LOINC IDs we filtered lab tests for HbA1c and glucose levels, identifying diabetic cases by two separate tests within two years, recording the date of the first. Medications. We identified diabetic patients through prescriptions for specific DM medications or non-specific medications when accompanied by a relevant diagnosis or lab test within two years. Finally, we combined data from diagnoses, labs, and medications to form the DM cohort, marking the earliest event date as an estimate for the DM onset. If a patient included a DM diagnosis in their first encounter, we marked that date as the DM onset estimate.

Figure 2B shows the diagnosis and procedure codes used to identify the four major complications among the diabetes patient cohort, including retinopathy, kidney disease (nephropathy), nerve damage (neuropathy), and major adverse cardiovascular events (MACE). The MACE events are acute myocardial infarction, stroke, heart failure, and hospitalization for revascularization procedures. For any of the four

major complications identified, we recorded the date of the first recorded encounter for each complication in the EHR as the endpoint event date.

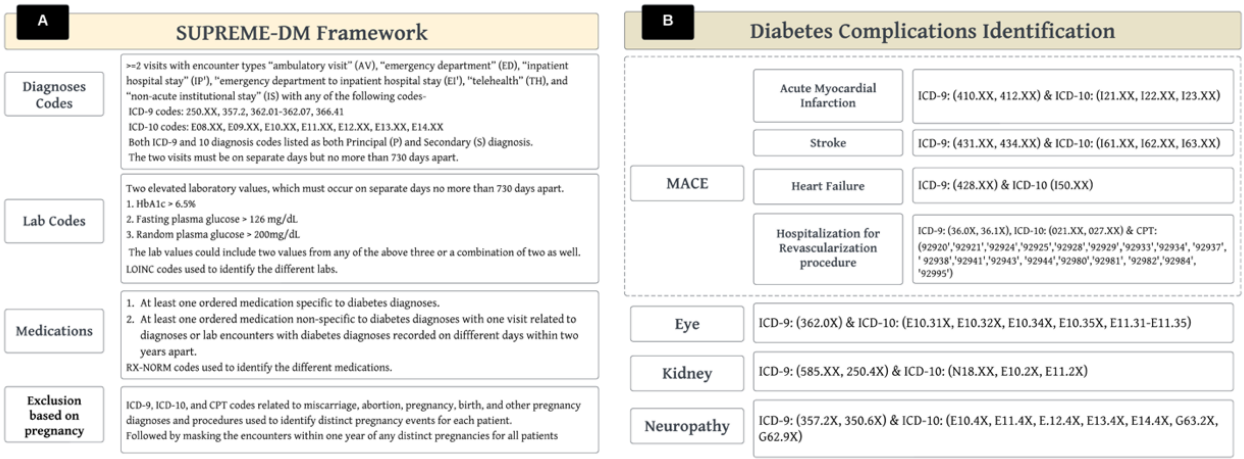


Figure 2 (A) shows the framework for constructing our diabetes cohort using SUPREME-DM from the MU EHR database. The process includes the exclusion of pregnancy-related encounters, identification of diabetes via ICD-coded visits and lab tests for HbA1c and glucose, and diabetic classification through specific medication prescriptions. The earliest diabetes indicator among diagnoses, lab results, or medications is marked as the estimated onset date for diabetes. Figure 2(B) illustrates the methodology for identifying major diabetic complications within the patient cohort. This includes the utilization of specific diagnosis and procedure codes to identify occurrences of retinopathy, chronic kidney disease, neuropathy, and MACE, with the latter encompassing myocardial infarction, stroke, heart failure, and revascularization hospitalizations. The initial encounter date for each complication is captured and recorded as the endpoint event date in the EHR data.

Outcomes, features, and study timeline. Figure 3-[1] shows how we structured a learning period for a hypothetical patient timeline in our EHR. This learning period determines the cutoff point for each patient in the identified DM cohort. The learning period for patients with any of the four diabetic complications spanned from their first EHR encounter up to the visit immediately preceding their first recorded complication event. For those without complications, the learning period extended to their last EHR encounter i.e. their learning period includes the entire EHR sequence. We also excluded patients whose diabetes diagnosis was recorded after their complication diagnosis to ensure diabetes was reported before the complication occurrences in the EHR timeline. The patient cohort derived from this process (cohort A) was used for pretraining. For finetuning and our Top-BERT experiments, we refined the cohort (cohort B) to only include patients with at least 5 but no more than 100 recorded encounters on different dates and with a minimum of five different diagnosis codes.

We defined the incidence for each complication (CKD, MACE, NEUR, and RET) within the first 12 months from the end of the learning period (prediction window). For instance, if a patient is diagnosed with MACE within the next 12 months, as shown in Figure 3-[1], after the end of its learning period, then the outcome is 1 else 0. Each patient can have multiple complications diagnosed in the prediction window. Thus, in the DM cohort, each unique patient ID is linked to four distinct labels, each binary and not mutually exclusive, to be used for our sequence classification task. For the binary classification task, we identified a common medical issue—prolonged hospital stays—assigning a value of 1 if the length of stay exceeded 7 days at any time during the EHR record of each patient. We extracted the labels of the four micro- and macro-vascular complications and prolonged hospital stay for all patients in the finetuning cohort B as our outcomes for the study. We curated separate datasets for pretraining and finetuning (cohorts A and B, respectively), extracting all encounters with diagnosis codes with admit-discharge dates and date of birth (DOB) within their specified learning period. The diagnosis ICD 9/10 codes are mapped to Phecodes[41] to reduce sequence dimensionality. The DOBs were used to calculate the age at each encounter date. Finally, we constructed a tabular representation of the EHR timeline for each patient, as shown in Figure 3-[2].

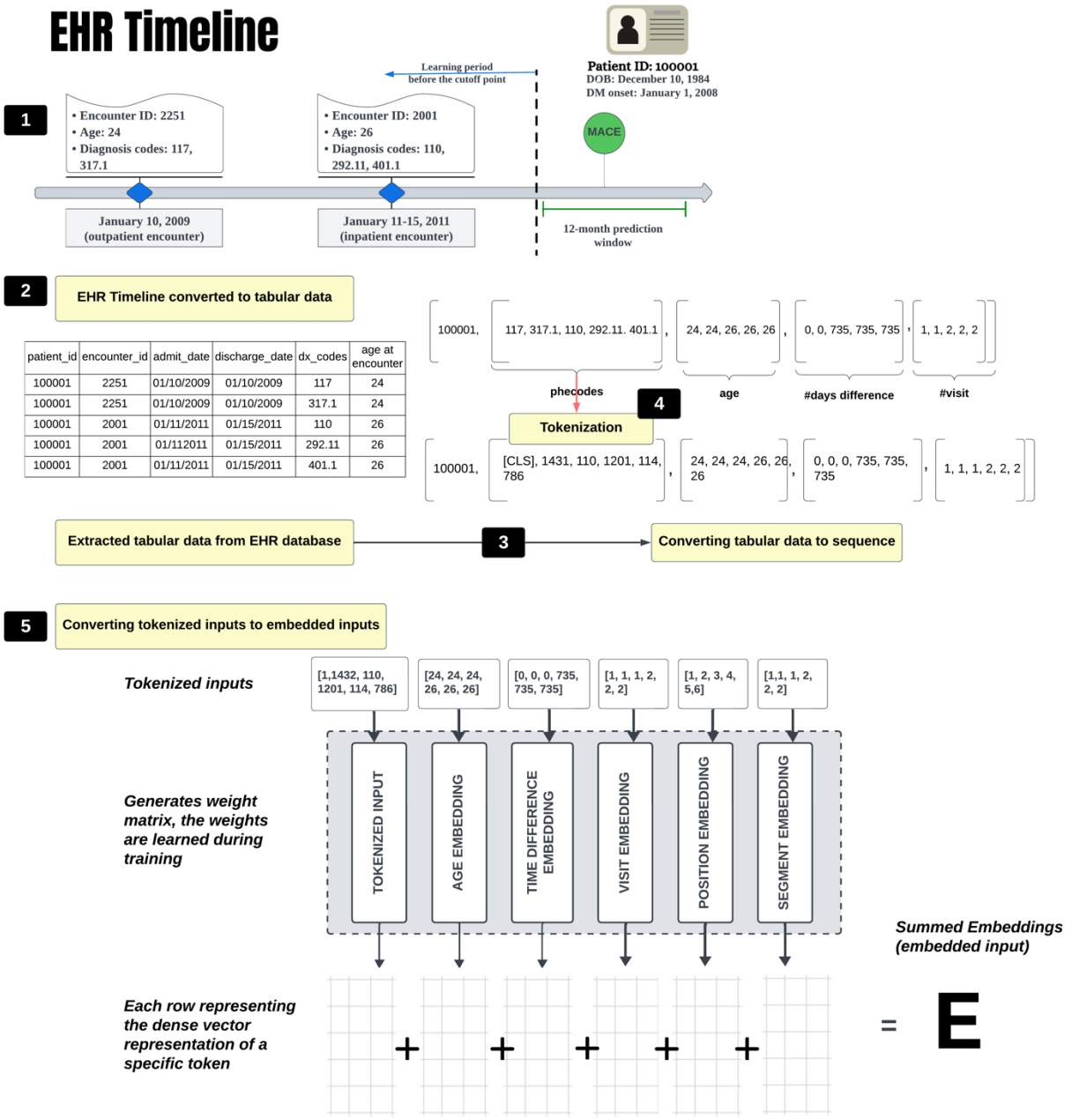


Figure 3. This figure illustrates [1] the methodology for structuring a learning period based on EHR data for patients with diabetes, defining the time frame leading up to either the earliest complication event or the last recorded encounter. The delineation of this period guided the selection of patient cohorts for pretraining and finetuning, with specific inclusion criteria based on encounter and diagnosis code counts to ensure relevance and accuracy in predictive modeling. The incidence of complications was identified in a designated prediction window following the learning period, contributing to the creation of outcome vectors for sequence classification tasks. This approach underpins the preprocessing of features for BERT, involving [2] the conversion of the timeline to structured tabular data, [3] the conversion of structured tabular data into sequential formats, and [4] the development of a tailored dictionary for EHR-specific tokenization,[5] culminating in the embedded input sequences that drive the training of the BERT-based model.

Feature preprocessing for BERT. In this study, we used both diagnoses lists and temporal features—age at visit, number of visits, and inter-visit time difference in days—as integral components of the input to the

embedding layer within our BERT-based model. To facilitate the training of BERT with EHR data, we initially converted the structured tabular data into a sequential format, as depicted in Figure 3-[3]. This entailed creating the temporal attributes as sequences, aligning each diagnosis code (in order of occurrence) to corresponding temporal attributes, and standardizing the sequence length for each patient's record. Acknowledging the distinct nature of EHR diagnosis codes as opposed to traditional text, we constructed a unique dictionary. This lexicon assigns a discrete numerical identifier to each diagnosis code and each special token, such as [CLS], [PAD], and [MASK], thus creating a bespoke mapping system tailored for EHR data. Subsequently, we employed this dictionary for tokenization, transforming the list of Phecodes into a sequence of tokens with a [CLS] token inserted at the beginning, as illustrated in Figure 3-[4]. These tokenized sequences are what we leveraged as the input features for the embedding layer. Figure 3-[5] shows the procedure for formulating the embedded input sequences essential for training our BERT-based model. Each tokenized input sequence with its corresponding temporal attributes is used as input for the embedding layer. This layer is responsible for transforming each token into a dense vector representation, generating weight matrices for each layer of the model. The dimensions of the matrices are determined by the sequence's length and hidden size parameter of the model. Finally, the summed embeddings are fed into the model's subsequent layer to complete the training process.

Experimental Design

Model training details and experiments. Our experimental design involved two distinct cohorts: an unlabeled cohort for pretraining (cohort A) and a labeled cohort (cohort B) for finetuning, as well as training our Top-BERT model. We randomly allocated cohort B into training, validation, and test sets in a 7:1:2 ratio for use in both finetuning and Top-BERT training. We conducted experiments to train, test, and evaluate the Top-BERT architecture on predicting four major micro and macro-vascular complications— CKD, MACE, NEUR, and RET— and prolonged hospital stays. Furthermore, we developed a pretrained model using cohort A and finetuned the model using cohort B to predict these micro and macro-vascular complications. We implemented a PyTorch[42] workflow to run our BERT-based experiments. Additionally, we benchmarked our model against the popular machine learning model XGBoost[43].

Top-BERT training and evaluation: We trained our Top-BERT model using three tasks simultaneously, namely, mask language model (MLM), binary classification for prolonged hospital stay, and a sequence classification for the 12-month prediction of the four major diabetic complications. For the MLM task, we masked 15% of the tokens in each input sequence, following a strategy where 80% of the masked tokens were replaced with a special [MASK] token (denoted as -1), 10% were replaced with a random token from the vocabulary, and the remaining 10% were left unchanged. The masked tokens were augmented with the processed input features to train the model. This masking strategy introduces noise and variability into the input data, encouraging the model to learn robust and context-dependent representations of the EHR data. The primary objective of the MLM task is to accurately predict the original tokens at the masked positions. During training, a cross-entropy loss function is utilized, which ignores the masked tokens to concentrate on the contextual learning of unmasked tokens. For evaluation, the model's performance is exclusively assessed on its capacity to predict the masked positions correctly.

Simultaneously, the binary task tokens were generated, indicating prolonged hospital stays (greater than 7 days) during the input feature-augmentation process for model training. For the prediction process, the logits from the model's output —derived from the pooled output and, subsequently, a dense layer— are processed through a softmax layer to calculate predictive probabilities. Binary cross-entropy loss function was used to minimize the prediction error. The third task, focused on sequence classification, predicts four types of micro and macro-vascular complications (CKD, MACE, NEUR, and RET). In this component, the model channels the pooled output through a dense layer to generate logits. These logits are then transformed into probabilities using the sigmoid activation function. The model is fine-tuned with a weighted binary cross-entropy loss function to optimize the predictive accuracy. The combined loss from all three tasks is

backpropagated through the layers to learn the weights. We used the AUROC metric to evaluate model performance as the primary evaluation metric like previously established frameworks[27,28]. We employed the best-performing trained model for evaluation against our test data and report all results on it.

Pretraining-finetuning: During the pretraining phase, we employed conventional MLM and binary tasks (prolonged hospital stay) utilizing cohort A. Post-training, the model demonstrating optimal performance was further finetuned on cohort B. Loading the pretrained weights, two finetuning variations were employed, each utilizing different encoder output strategies. Finetuning-A employed conventional [CLS] token embeddings. Finetuning-B innovated with a custom prediction head utilizing the last encoder layer's embeddings refined through three dense layers and ReLU activations, with the final layer producing logits. We implemented a weighted random sampler during batch training to address class imbalance in the four label predictions.

Comparison with an ML model: XGBoost was employed as a time and context unaware benchmark for comparison with our time and context aware BERT-based models. We transformed the diagnosis codes into one-hot encoded features across our training, validation, and test datasets. Each Phecode was denoted by a binary flag, and we also included the age at the last visit as an additional feature. We utilized a multioutput strategy with XGBoost to predict the four complications simultaneously.

Metrics for model comparison: To assess the performances of the four classification tasks (CKD, MACE, NEUR, and RET) in BERT-based models and XGBoost, we computed micro-averaged AUROC (mAUROC). The mAUROC pools the individual true positives, false positives, false negatives, and true negatives across all classes and then computes the AUROC from these combined totals, which indicates the overall performances of the models in distinguishing between both majority and minority classes across various thresholds[44]. Additionally, to gain insights on the differences and similarities of Top-BERT and XGBoost in handling class imbalance, we compared the model performance for each classification task using precision, recall, F1-score, Mathew's correlation coefficient (MCC) and confusion matrix computed at varying threshold values (10%, 30%, 50%, and 80%).

Temporal representation comparison using embeddings. In our investigation, we explored the impact of temporal factors—patient age, visit order, and inter-visit time differences—when integrated into the embedding process of predictive models for diabetic complications. We conducted ablation studies to comprehensively evaluate the utility of various temporal embeddings.

In our Top-BERT experiments, we compared 15 distinct models, each featuring a distinct embedding layer architecture that integrated various temporal components. The distinct models we compared also include the three pioneering frameworks of BERT in EHR data– BEHRT[21] and Med-BERT[27]. The mAUROC metric served as our primary performance indicator, reflecting each model's discriminative power in predicting diabetic complications. Concurrently, we utilized Shannon's entropy[32]—a measure of the unpredictability or complexity of information content—to gauge the informativeness of the embeddings generated by each model. We obtained the embeddings from the final encoder layer for each model, transformed them into probabilities via softmax, and then calculated their total entropy using the formula $H(X) = -\sum_{x \in X} p(x) \log p(x)$. A higher Shannon's entropy value indicates a richer, more complex embedding representation, suggesting that the model captures a greater amount of information from the input data. Information gain was calculated by comparing the entropy values of each model to that of a comprehensive model inclusive of all temporal factors— incorporating input, visit number, inter-visit time difference, age, and positional and segmental embeddings—providing a metric for the relative improvement in predictive power.

Furthermore, we performed ablation studies across different embedding layer architectures within our pretraining and finetuning workflow, creating 15 unique pretrained models. Subsequently, we executed 15

finetuning experiments each for scenarios A and B. By juxtaposing mAUROC scores with information gain values; we aimed not only to identify the model with the highest predictive accuracy but also to discern which embedding design best encapsulates the complexity of patient history in our single-center dataset. This dual assessment allowed us to balance the trade-off between model simplicity and the depth of temporal understanding necessary to accurately depict the progression towards diabetic complications.

Model Explanations. We have tailored the Integrated Gradients[45] methodology for application on our Top-BERT model to discern feature-level attributions that influence our predictive models on both an individual patient and a global level. Integrated Gradients is a feature attribution method that assigns an importance score to each input feature of a neural network by integrating the gradients of the model's output with respect to the input features, tracing a path from a given baseline to the actual input. This method is designed to satisfy two fundamental axioms: sensitivity and implementation invariance, ensuring reliable and consistent attributions.

Formally, for a given input x and a baseline x', the integrated gradient along the i-th dimension is defined as:

$$\text{IntegratedGradients}_i(x) = (x_i - x_i') \times \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha \times (x - x'))}{\partial x_i} \, d\alpha$$

Here, $F: \mathbb{R}^n \rightarrow [0,1]$ is the model, x is the input, x' is the baseline, x_i is the i-th feature of x, and $\frac{\partial F}{\partial x_i}$ is the gradient of F with respect to x_i. The integral accumulates the gradient for feature i at all points interpolated between the baseline and the input. A practical baseline for language models is the all-zero input embedding vector. We approximate the integral via a summation called the Reimann approximation. To apply Integrated Gradients to Top-BERT, we consider the model function $F(x)$ as the output logit resulting from the input sequence x after transformation through the model's embedding and subsequent BERT layers: $F(x) \rightarrow Head\left(BertLayers\left(EmbeddingLayer(x)\right)\right)$, which is $\boldsymbol{h}_{seq}$ for Top-BERT. We divide the linear path between the baseline and the input into α increments to calculate the gradients at each step, which are then aggregated and normalized to determine the contribution of each feature to the prediction for each outcome. Additionally, to identify the global-level feature attributions, we averaged attribution scores for each feature in each patient and then calculated a sum over these feature-level attributions across all patients for each outcome. Furthermore, we computed mean attribution scores within distinct age groups to identify age-specific patterns of feature importance for each outcome, thereby enhancing the explainability and transparency of our model in the clinical domain.

Algorithm 1. Adaptation of Integral Gradients Algorithm for Sequence Classification using Top-BERT Input:
M: Top-BERT model trained for the sequence classification task,
N: number of classification tasks,
$Embedding$: embedding layer of model M,
$BertLayers$: encoder layers of model M
$\boldsymbol{h}_M$: prediction head for sequence classification in model M,
F_k: output logits for k^{th} outcome of the classification tasks,
$X^{(l)}$: input sequence of length l for M,
$X'^{(l)}$: baseline input sequence of length l,
α: number of steps for Reimann
Output:
IG_k: Attribution Sequence of length l for k^{th} Outcome

Load Model M
IntegratedGradient:
  for $k = 1$ to N:
    Initialize attribution sequence IG_k to zero vector of length l

for each token position $i = 1$ to l:
for steps $= 1$ to α :
$\alpha_{step} = \text{step} / \alpha$
$X_{\text{interpolated}} = X'^{(l)} + \alpha_{step} \times (X^{(l)} - X'^{(l)})$
$E_{\text{interpolated}} = Embedding(X_{interpolated})$
$F_{k,\text{interpolated}} \rightarrow \boldsymbol{h}_M(BertLayers(E_{interpolated}))$
$\text{gradient} = \frac{\partial F_{k,\text{interpolated}}}{\partial X_{\text{interpolated}}}$ at position i
$IG_k[i] \pm \left(X^{(l)}[i] - X'^{(l)}[i]\right) \times \text{gradient} \times \frac{1}{\alpha}$
end for
Normalize $IG_k[i]$ by dividing by the sum of $IG_k[i]$ across all α steps
end for
output IG_k
end for
end IntegratedGradient

Results

Data summary

Figure 4 provides a detailed demographic and clinical profile of the diabetes patient cohort used in this study. Table 4A contrasts two subsets: the initial pretraining group (Cohort A), comprising 50,993 patients, and the subsequent finetuning/Top-BERT group (Cohort B), which includes 36,539 patients. The average number of visits per patient for pretraining cohort A (~31 visits) is higher than cohort B (~27 visits). Both subsets exhibit similar averages in number of diagnoses per patient (~44 diagnoses per patient). The average age for both cohorts is about 60 years and primarily white (~89% white), the cohorts maintain a consistent gender ratio. Prolonged length of stay was approximately 19.4% for both cohorts. Figures 4B through 4E show further insights into Cohort B, shedding light on healthcare interaction patterns and the prevalence of specific health complications within this group. Specifically, Table 4D indicates high occurrences of hypertension (69%), hyperlipidemia (51.7%), obesity (34.1%), and tobacco use disorder (33.53%) within the cohort, which are predominantly categorized under endocrine/metabolic and circulatory system

disorders. Furthermore, approximately 10.7% of patients had neuropathy, 10.9% had MACE, 7.4% had CKD, and 2.7% had retinopathy.

Table A. Cohort characteristics for this study

Characteristic	Pretraining Cohort (A)	Finetuning/ top-BERT Cohort (B)
Cohort Size	50,993	36,539
Cohort Size * Total Number of Visit for Each Patient	6,970,867	4,427,406
Average Number of Visits Per Patient (Median)	31.23 (14.0)	26.79 (18.00)
Average Number of Codes Per Patient (Median)	43.5 (31.0)	43.42 (36.0)
Average Maximum Age of Each Patient (Median)	60.34 years (61.62 years)	60.38 years (61.59 years)
Gender, Male (%)	24, 664 (48.37%)	17,386 (47.58%)
Race, White (%)	44,965 (88.18%)	32,432 (88.76%)
Race, Black (%)	3,748 (7.35%)	2,700 (7.39%)
Vocab Size	1861	1861
Prolonged Length of Stay (LOS, %)	9,877 (19.37%)	7,079 (19.37%)

Figure B. Distribution of number of visits per patient for Cohort B

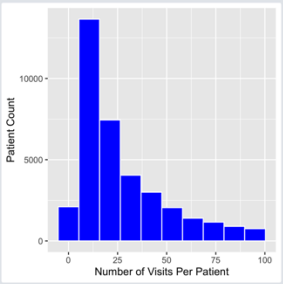


Figure C. Distribution of number of diagnoses per patient for Cohort B

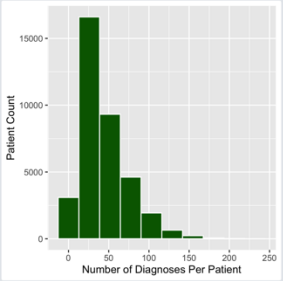


Table D. Highly occurring diagnoses in Cohort B

Phenotype	Category	Count	(%)
Type 2 diabetes	endocrine/metabolic	26502	(72.53)
Essential hypertension	circulatory system	25196	(68.96)
Hyperlipidemia	endocrine/metabolic	18882	(51.68)
Pain in joint	musculoskeletal	16178	(44.28)
Obesity	endocrine/metabolic	12452	(34.08)
Tobacco use disorder	mental disorders	12253	(33.53)
Abdominal pain	symptoms	12191	(33.36)
Back pain	symptoms	11912	(32.60)
Disorders of lipoid metabolism	endocrine/metabolic	11137	(30.48)
Abnormal electrocardiogram [ECG] [EKG]	circulatory system	10785	(29.52)
Malaise and fatigue	symptoms	10605	(29.02)
GERD	digestive	10438	(28.57)
Nonspecific chest pain	circulatory system	9656	(26.43)

Figure E. Distribution of the four major vascular complications in Cohort B

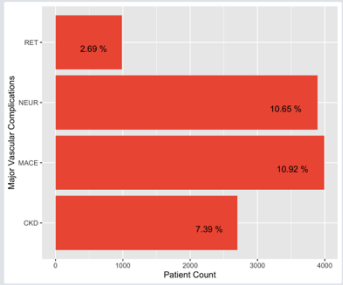


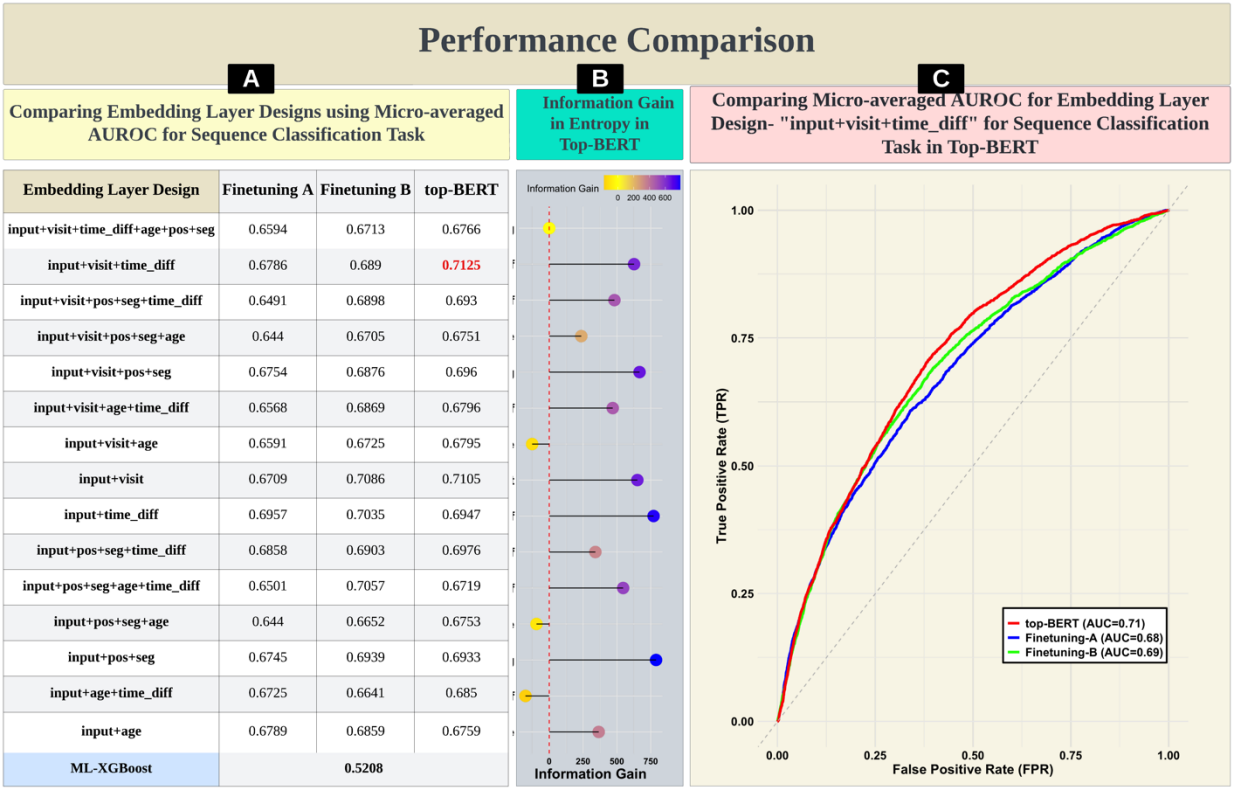
Figure 4: This figure provides a comprehensive summary of the cohort's demographic and health status, the distribution of healthcare interactions, and the prevalence of significant health conditions, which are critical for the subsequent analysis of the study. Table 4A outlines the characteristics of the study cohorts, revealing a pretraining cohort (A) size of 50,993 and a finetuning/Top-BERT cohort (B) size of 36,539. Figure B illustrates the distribution of the number of visits per patient for cohort 4B, which shows a right-skewed distribution, indicating that most patients have fewer visits, with the number tapering off as the visit number increases. Figure C depicts the distribution of the number of diagnoses per patient for cohort B, which is also right-skewed, with more patients having diagnoses number in the mid-range of 100. Table 4D lists the most frequently occurring diagnoses in cohort B, with type 2 diabetes, essential hypertension, and hyperlipidemia being the most common. These conditions are predominantly categorized under endocrine/metabolic and circulatory system disorders, reflecting the health concerns prevalent in the cohort. Figure E depicts the relative distribution of the four primary study outcomes, indicating a higher frequency of neuropathy and major adverse cardiovascular events (MACE) within the patient cohort.

Model Performance

In evaluating the performances of the three distinct architectures for predicting four major complications within Cohort B, our Top-BERT model outperformed the finetuning adaptations of pretrained models and the time and context unaware approach using XGBoost. Our test data consisted of 7.25% of CKD labels, 11.6% of MACE, 10.9% of NEUR, and 2.7% of RET. As shown in Figure 5A, the highest mAUC was achieved by the Top-BERT design input+visit+time_diff (mAUC of 0.7125), while the same embedding configuration in the finetuning experiments had lower mAUCs (0.6786 and 0.689, respectively). In contrast, XGBoost yielded an mAUC of 0.5208. Moreover, Figure 5C highlights the ROC of the best-performing model configuration (input+visit+time_diff) with the finetuning frameworks at various thresholds.

We observed (Figure 5B) that an embedding incorporating input features with temporal elements, such as visit number and inter-visit time difference, alongside positional and segmental embeddings, yielded a higher information gain than the more complex, feature-rich embedding design (input+visit+time_diff+age+pos+seg). Within Top-BERT, the embedding approach utilized in BEHRT (input+age+pos+seg)[21] demonstrated a lower information gain, whereas the Med-BERT (input+visit)[27] embedding configuration showed higher information gain, while with AUCs of 0.6753 and 0.7105, respectively.

Table S2 in the supplementary document compares the five Top-BERT embedding configurations (having comparable mAUROCs) and XGBoost using various metrics at varying thresholds to gain insights on their differences and similarities in handling class imbalance. We observed that higher precision values and lower recall values obtained by XGBoost were consequences of conservatively predicting fewer number of positive cases (both true positives and false positives) in all classification tasks. On the other hand, Top-BERT showed better measure of separability, as indicated by higher mAUROC values, and higher efficiency in accurately predicting the positive classes compared to XGBoost (details in supplementary document).



Embedding Layer Design	Finetuning A	Finetuning B	top-BERT
input+visit+time_diff+age+pos+seg	0.6594	0.6713	0.6766
input+visit+time_diff	0.6786	0.689	0.7125
input+visit+pos+seg+time_diff	0.6491	0.6898	0.693
input+visit+pos+seg+age	0.644	0.6705	0.6751
input+visit+pos+seg	0.6754	0.6876	0.696
input+visit+age+time_diff	0.6568	0.6869	0.6796
input+visit+age	0.6591	0.6725	0.6795
input+visit	0.6709	0.7086	0.7105
input+time_diff	0.6957	0.7035	0.6947
input+pos+seg+time_diff	0.6858	0.6903	0.6976
input+pos+seg+age+time_diff	0.6501	0.7057	0.6719
input+pos+seg+age	0.644	0.6652	0.6753
input+pos+seg	0.6745	0.6939	0.6933
input+age+time_diff	0.6725	0.6641	0.685
input+age	0.6789	0.6859	0.6759
ML-XGBoost	0.5208		

Figure 5. This figure shows a comprehensive depiction of the performance comparison of our proposed Top-BERT framework with the two finetuning frameworks adapted for this study for predicting the four major complications within Cohort B. Figure 5A displays a table comparing various embedding layer designs based on micro-averaged AUROC scores for three different model training scenarios: Finetuning A, Finetuning B, and the final Top-BERT model. The highest-performing design in the Top-BERT model is highlighted, suggesting it has the best trade-off between complexity and performance. Figure 5B visualizes the information gain in entropy for the Top-BERT model across different embedding designs. Each point represents a different design, with its position on the x-axis indicating the amount of information gain using the most complex model (input+visit+time_diff+age+pos+seg) as a baseline for comparing the gain. This graph helps assess which design captures the most relevant information from the data. Figure 5C compares the micro-averaged AUROC curves of the Top-BERT model and the two finetuning phases, illustrating the true positive rate (TPR) against the false positive rate (FPR) for highest-performing design. The area under the curve (AUC) for each model is annotated, allowing for a direct comparison of their predictive performance.

Model Explanations

Figure 6 highlights the aggregated global-level feature importance for the four outcomes of this study: CKD, MACE, NEUR, and RET. In Figure 6, the left panel shows the common diagnoses predictive of all four outcomes, while the right panel shows the significant features contributing (positively) to each outcome separately. In the figure, the color intensity representing the attribution score for each diagnosis reflects the combined influence of the average attribution score per patient and the prevalence of that diagnosis within each respective outcome cohort.

Common Predictive Features

Metabolic abnormalities such as vitamin-D deficiency, hypopotassemia, magnesium metabolism disorders, and disorders of fluid, electrolyte, and acid-base balance emerged as common important predictors for all studied outcomes. Additionally, circulatory system conditions, including hypertension, tachycardia, and peripheral vascular disease, alongside respiratory system disorders and hematopoietic conditions, such as chronic anemia, were found as shared predictive features for increased risk of all four outcomes (Figure 6, left panel).

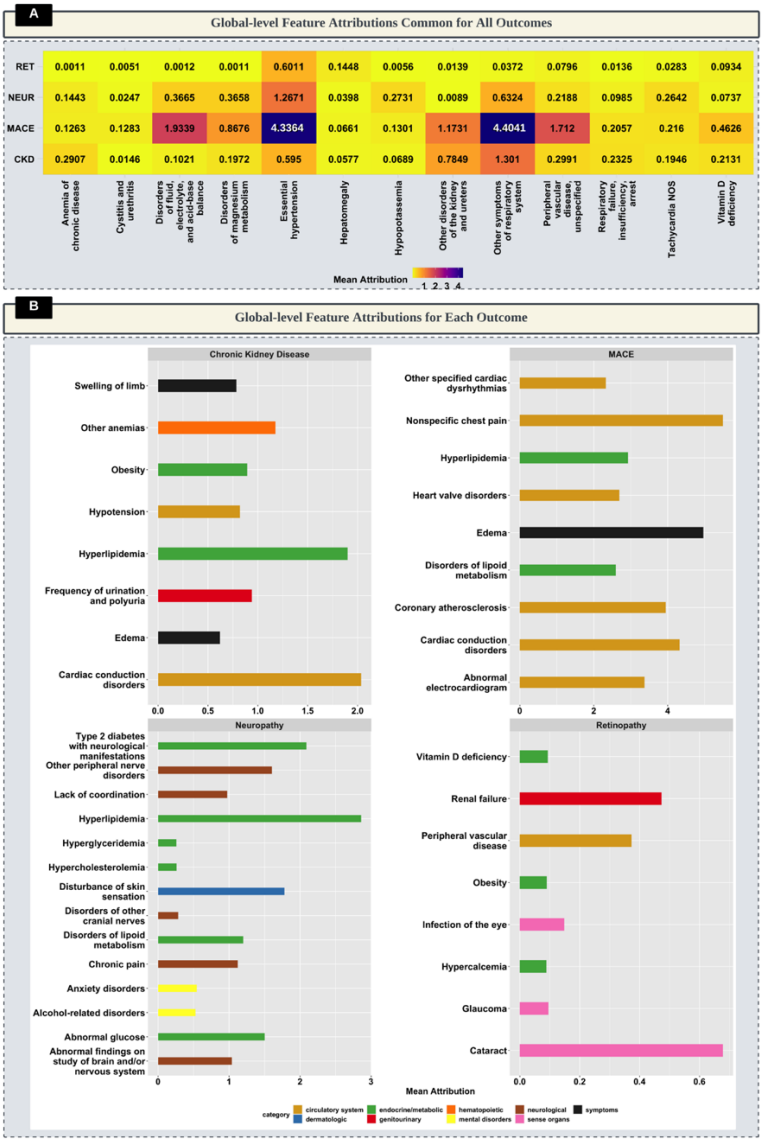


Figure 6. Global-level feature attributions for four major diabetes-related complications. The top panel (A) shows the common feature attributions across all outcomes, while the bottom panel (B) details the top contributing features for each specific complication: chronic kidney disease (CKD), Major Adverse Cardiac Events (MACE), Neuropathy (NEUR), and Retinopathy (RET). In Figure 6A, each column represents a diagnosis with its corresponding aggregated mean attribution score, represented in color-coded heatmaps, indicating the strength of prediction with the complication outcomes. In Figure 6B, each row is the aggregated mean scores for the diagnosis and categories of diagnoses are color-coded, as indicated in the legend, facilitating a comparative visualization of feature importance across the different diagnostic classifications. Note that the mean attribution scales vary for each complication to optimize the visual representation of the data, facilitating the comparison of influential predictive diagnoses within each respective outcome.

Outcome-Specific Feature Importance

Edema and cardiac conduction disorders highly indicated increased risk for CKD and MACE. Similarly, disorders of lipoid metabolism had high attribution for predicting both MACE and NEUR. Hyperlipidemia had high attribution scores for CKD, MACE, and NEUR. Additionally, obesity was highly predictive for both CKD and RET.

Symptoms such as swelling of limbs, low blood pressure (hypotension), and urinary system conditions such as disorders of kidney/ureter and frequency of urination were found to be notable for their contribution to CKD risk. For MACE, various circulatory system disorders such as abnormal electrocardiogram, coronary atherosclerosis, chest pain, cardiac dysrhythmias, and heart valve disorders– were identified as critical indicators. NEUR risk was closely linked with early neurological manifestations from Type 2 diabetes, peripheral nerve disorders, coordination issues, and chronic pain accompanied by abnormal findings related to the brain or nervous system. Interestingly, mental health disorders, including anxiety and alcohol-related disorders, also had high attribution for predicting NEUR. Furthermore, metabolic conditions such as hypercalcemia with early onset of eye-related symptoms such as infection, glaucoma, cataracts attributed significantly to increased risk of RET.

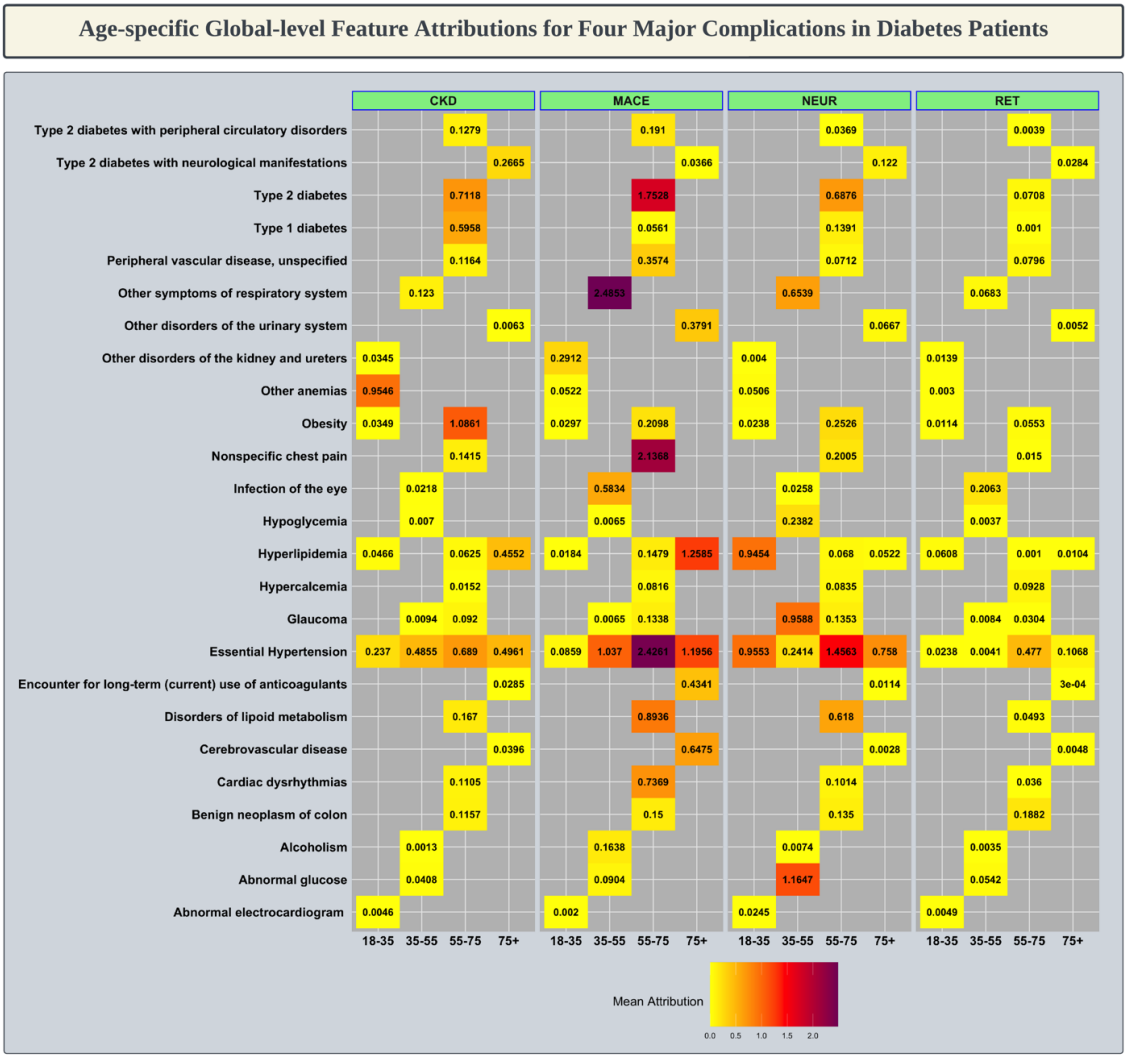


Figure 7. Heatmap illustrating age-specific global-level attributions of the common predictive features for four major complications in diabetes patients (each diagnosis is present in at least one age group). Each row represents a distinct diagnosis, while columns

denote the patient age groups and the complications—chronic kidney disease (CKD), major adverse cardiovascular events (MACE), neuropathy (NEUR), and retinopathy (RET). Color intensity reflects the mean attribution value for each feature, with warmer colors indicating higher attribution (with higher incidence) and cooler colors indicating lower attribution, signifying the relative importance of each feature in the model's predictions across different age brackets.

Age-specific Feature Importance

Figure 7 shows the age-specific predictive features common across all outcomes (each diagnosis was present in at least one age group) of this study. Hypertension was shown to be a predictive factor for all outcomes for all age groups. In the 18-35 age group, the presence of hypertension, hyperlipidemia, and obesity attributed to an increased risk for all four outcomes. For individuals aged 35-55, abnormal glucose levels, alcoholism, hypoglycemia, and respiratory disorders were highlighted as significant predictive indicators. Among those in the 55-75 age range, obesity, disorders of lipoid metabolism, and diabetes diagnoses stood out as key common predictors of high risk for all outcomes. In the population over 75, hyperlipidemia and the prolonged use of anticoagulants were identified as common indicators across the studied outcomes.

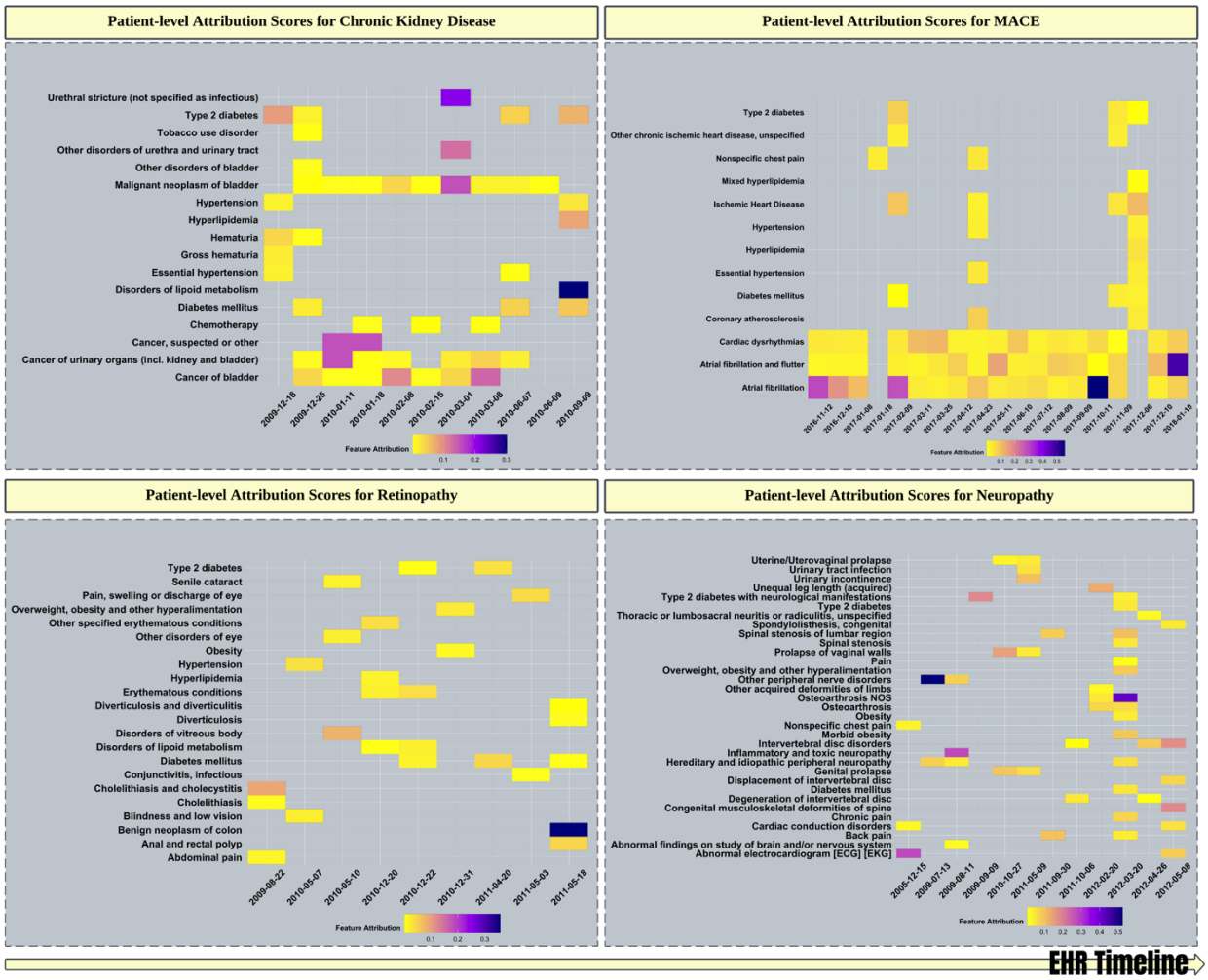


Figure 8. Heatmaps of patient-level attribution scores for four randomly chosen patients within our study cohort who developed major diabetic complications within a year following their last recorded visit. The visualizations present individualized feature attributions within each patient's EHR timeline for chronic kidney disease (CKD), major adverse cardiovascular events (MACE), retinopathy, and neuropathy. Each row represents a specific diagnosis, while columns represent the encounter dates, showcasing how attribution scores vary throughout the patient's medical history. Color gradients indicate the magnitude of the feature's

attribution to the respective complication, with warmer colors denoting higher attribution scores. This patient-centric analysis highlights the differential and time-related impact of various medical conditions on the risk of each complication.

Patient-specific Feature Importance

Figure 8 illustrates the attribution scores of various conditions diagnosed, computed using IG, for each visit recorded in the EHR. The figure provides a visual narrative of patient-specific features contributing to increased risk for the four outcomes examined, as determined for four randomly selected patients. For example, the timeline for a patient at elevated risk for Major Adverse Cardiac Events (MACE) (top-right quadrant) reveals atrial fibrillation and cardiac dysrhythmia as consistent predictive features across visits from November 2016 to January 2018, preceding a MACE diagnosis within the subsequent year. Notably, hyperlipidemia and hypertension emerged as indicative features of MACE risk in later visits. In another case, a patient at high risk for chronic kidney disease (CKD) (top-left quadrant) showed predictive features such as kidney or bladder cancer, along with hypertension, hyperlipidemia, and disorders of lipoid metabolism in visits leading up to the CKD diagnosis.

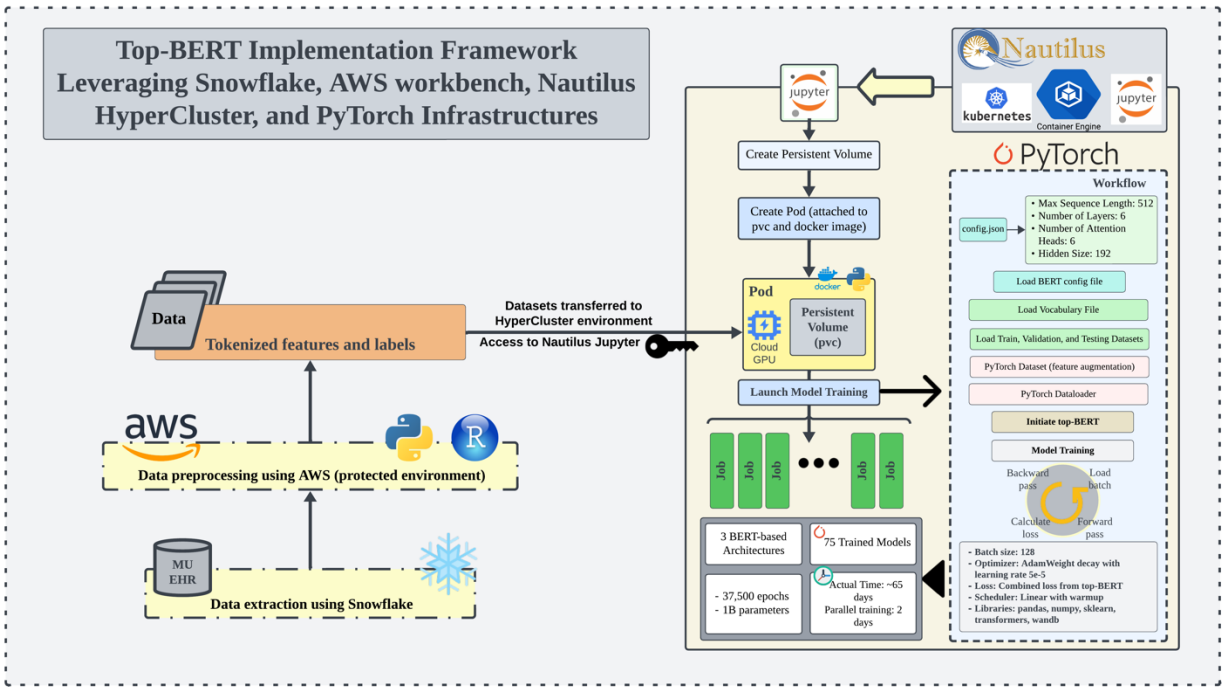


Figure 9. The diagram outlines the Top-BERT Implementation Framework, integrating various technological platforms for model development and training. The process begins with data extraction from MU EHR using Snowflake, then preprocessing in AWS's secure environment. Tokenized features and labels are then transferred to the HyperCluster environment facilitated by Nautilus Jupyter to access GPU resources. Model training is initiated in this environment via persistent volumes and pods, leveraging Kubernetes orchestration. The PyTorch workflow incorporates configuration of BERT's parameters, vocabulary loading, and dataset preparation for training 75 models across 3 BERT-based architectures. Training efficiency is exemplified by the reduction of actual training time to approximately 65 days, achieved in just 2 days due to parallel computing. This streamlined process underpins our study's robust analytical capability.

Implementation Details

Figure 9 outlines the computational framework employed in our study, utilizing Snowflake, AWS Workbench, the Nautilus HyperCluster, and PyTorch(2.0). Our process began with executing SQL queries within Snowflake for cohort identification from our EHR database. Subsequent data preprocessing and feature generation for model training were conducted within AWS's secure environment.

The tokenized features and labels were transferred to the HyperCluster environment called NRP (National Research Platform) Nautilus, a nationwide cyberinfrastructure led by the Greater Plains Network. Figure 9

further illustrates the use of Jupyter IDE to implement model training across multiple experiments. We allocated persistent volumes to store datasets and orchestrated containerized jobs with the necessary docker image and GPU support. Specifically, NVIDIA A10 GPUs with 25 GB each were employed to facilitate parallel execution of tasks. Figure 9 also presents the workflow of our PyTorch-based implementation. It details the hyperparameters selected for our experiments and lays out the sequential steps in training the models. This systematic approach underpins the rigorous development and validation of our predictive models. We customized the Transformers library's BERT model for model training to construct our unique Top-BERT architecture. Monitoring and managing training progress were achieved through Weights and Biases (wandb.ai). Using Nautilus' parallel computing capabilities drastically reduced our total training duration for all experiments from an estimated 65 days to just 2 days, leading to significant efficiency gains.

Additionally, we implemented XGBoost to predict four micro and macro-vascular complications, utilizing the xgboost library, scikit learn's MultiOutputClassifier, and the Optuna library for hyperparameter optimization. Model evaluation metrics were computed using scikit learn.

Discussion

Methodological and Technical Advancement

Our study demonstrates a notable advancement in the application of BERT's architecture for clinical predictions from EHR data, particularly when constrained by the sample size typical of single-center datasets. Through the innovative implementation of task-oriented predictive (top)-BERT, we have demonstrated the adaptability and strength of the BERT architecture in facilitating an end-to-end training and evaluation approach. Top-BERT, utilizing the sequential input structure, embedding layer, and encoder stacks inherent to BERT adopts a multitask approach that integrates the conventional Masked Language Model (MLM), a binary classification for prolonged hospital stays, and multilabel sequence classification for micro and macro-vascular complications in diabetic patients. Our findings showed that Top-BERT can outperform both standard pretraining-finetuning BERT applications and traditional machine learning models such as XGBoost.

The inclusion of a binary classification task for prolonged hospital stays was inspired by Med-BERT (2021)[27] suggestion to substitute the generic Next Sentence Prediction (NSP) task with more contextually relevant tasks for EHR data. Moreover, in our approach, we adopted the strategy of including a [CLS] token—following the work of Li et al. (2020)[21] —at the beginning of each input sequence, leveraging its role as a summarizing representation of the sequence. The [CLS] token's aggregated representation provides a distilled feature vector that encapsulates the contextual information from the entire sequence, which is pivotal for downstream classification tasks. Additionally, the differing embedding layer designs in earlier BERT models for EHR data emphasize the necessity to investigate how these designs affect the model's ability to accurately represent EHR temporality. By conducting ablation experiments, we evaluated the impact of integrating temporal factors—patient age, visit sequence, patient age at encounter, and inter-visit durations —on the predictive accuracy for diabetic complications.

We combined AUROC scores with Shannon's information entropy to evaluate model performance, optimizing a balance between model simplicity and performance accuracy. We found that simpler temporal embeddings, such as visit number and inter-visit durations, offered an optimal trade-off, achieving high predictive accuracy, which is likely attributable to the information gain rooted in its representation. Combining input with visit number and inter-visit durations achieved the highest micro-averaged AUROC of 0.713 in predicting the four outcomes. This represents an improvement of 4.8% and 3.3% over the respective finetuning frameworks and a substantial 27.0% over the AUROC achieved by XGBoost. Notably, within Top-BERT, this embedding combination surpassed the AUROCs of BEHRT's and Med-BERT's distinct embedding designs by 5.22% and 0.28%, respectively. Top-BERT also showed improved

performance over the finetuning frameworks of BEHRT and Med-BERT (enhancing AUCs by margins of 9.6% & 6.6% and 5.83% & 0.55%, respectively). Furthermore, Top-BERT showed consistent capability in handling the class imbalance with higher predictive true positive rates than XGBoost- further underscoring the potential of Top-BERT in discriminating between clinical tasks, especially in limited sample cases.

The combined use of self-supervised and supervised learning methods within Top-BERT utilizes the distinct benefits of multitask approach. Unsupervised learning methods identify underlying patterns in large volumes of unlabeled data, while supervised learning refines this understanding by directing the model's focus toward accurately predicting designated clinical outcomes. This approach gives Top-BERT the unique ability to derive generalized insights from data and perform clinical task predictions independent of any pre-established knowledge base from pretrained models. Multitask learning has been shown to obtain a more robust shared representation of the tasks that effectively can mitigate the sparsity of labeled data, enhancing model performance, faster model convergence, and reducing overfitting risks[46]. This accounts for the case that Top-BERT outperformed the two finetuning frameworks without relying on class balancing techniques to address label sparsity in the dataset, whereas the finetuning frameworks required weighted batch samplers to improve their learning. Moreover, Top-BERT's unified multitask learning approach achieves time efficiency and optimized performance, reaching convergence within 350 epochs (20-24 hours of training), compared to the pretraining-finetuning framework to our single-center EHR data which required approximately 28-30 hours for pretraining (epochs = 500) and an additional 8-10 hours for fine-tuning. This increase in efficiency, especially with limited data, further underscores the effectiveness of our approach.

Common Predictive Risk Factors and Clinical Relevance

The insights from our model explanations reveal the significant predictive diagnoses used by our model to make predictions for each diabetes-related complications, including chronic kidney disease, major adverse cardiac events (MACE), retinopathy, and neuropathy. Hypertension, prevalent in approximately 69% of our cohort, was found to be a key predictive factor for all studied outcomes across all age groups (Figures 6 and 7). Our findings resonate with established clinical evidence linking hypertension to an escalated risk of diabetic complications[47,48]. Furthermore, hypertension with diabetes is associated with a 6-fold increase in the risk of cardiovascular events, a risk that escalates further with the coexistence of chronic kidney diseases[48]. Additionally, our study identified hyperlipidemia and obesity as contributory to organ damage––reinforcing their role in the pathogenesis of diabetes-related complications[49].

Metabolic imbalances such electrolyte balance disorders—specifically potassium, magnesium, and phosphate—emerged as significant predictors across all complications, reflective of the complex interplay in diabetes management (Figure 6)[50,51]. Diabetes-related electrolyte imbalances stem from renal issues, absorption problems, acid-base imbalances, and extensive medication use. Low serum magnesium is linked to key diabetes complications, including retinopathy and heart disease, while polypharmacy can lead to hypopotassemia, further increasing cardiovascular risk. Our findings also showed association of vitamin D with both micro and macro-vascular complications, as well as identified anemia a significant predictive factor, corroborating existing clinical evidence[51, 52].

The age-specific patterns of symptoms suggest that for younger individuals, addressing modifiable risk factors such as hyperlipidemia and obesity could be crucial in reducing the risk of diabetes-related complications. Tailoring interventions for middle-aged individuals by monitoring glucose and addressing lifestyle factors like alcohol consumption while managing lipid levels in older adults may be vital. Thus, the common risk factors identified in this study highlight the significance of regular monitoring of glucose level, blood pressure, serum electrolytes and vitamin levels, hemoglobin, lipid profile, and weight in diabetes patients to improve their overall healthcare outcomes and hence prevent the onset of other complications[53,54].

Symptoms of Complications and Clinical Implications

Our findings delineate a trajectory of early clinical manifestations detectable in EHR data associated with an increasing risk of the studied complications. For example, patients at elevated risk for chronic kidney disease had earlier encounters indicating potential kidney dysfunction, including symptoms like limb swelling, increased frequency of urination, and fluctuating blood pressure. Patients at increased risk for cardiovascular complications presented with symptoms such as abnormal electrocardiogram readings and various cardiac anomalies, while those at risk for neuropathy exhibited chronic pain alongside other neurological symptoms. Similarly, individuals facing a risk of retinopathy had historical clinical encounters related to eye conditions. These insights highlight the essential role of proactive symptom assessment and the management of comorbid conditions in preventing the advancement of complications in patients with diabetes, emphasizing the medical necessity for comprehensive evaluation as a fundamental element of preventive care strategies[54].

Limitations and Future Goals

Although our study achieved considerable progress in predicting clinical outcomes from EHR data enhancing BERT framework, we recognize several limitations that open opportunities for future research.

Integration of Numerical Features: Our future goal is to improve the model performance in predicting the diabetic complications, especially leading to higher AUC values. Prior studies have shown to improve model performance with inclusion of numerical features such labs and vitals (body mass index, blood pressure, lipid profiles, etc.)[15,55,56]. BERT inherently processes inputs as tokens, including numerical features. This presents a limitation as it does not fully exploit the quantitative nature of these features. Thus, we will focus on integrating numerical data, such as lab values and vital signs, more effectively within our model architecture, which may potentially lead to enhanced model performance and identification of modifiable predictive factors.

Model Fairness Studies: Our research also highlights the role of thorough documentation and patterns of healthcare utilization in the EHR, which significantly contribute to the predictive features of our models. For example, patients with higher healthcare utilization tend to accumulate more diagnostic entries in their EHRs, influencing the model's predictions. As AI and machine learning become more prevalent in healthcare, it's critical to maintain model integrity and prevent biases that could adversely affect certain groups of patients. In our future work, we aim to conduct thorough model fairness studies to identify and mitigate any biases and confounding factors, ensuring equitable predictions across all demographics.

Multi-center Data Expansion: The current study's findings are based on single-center data, which may not capture the diversity of patient populations and practice patterns. Expanding the dataset to include multiple centers will allow our model to learn from a broader range of patient encounters, enhancing its generalizability and robustness across different clinical environments.

Thus, addressing these limitations is pivotal for advancing our research. The envisioned improvements and expansions will aim not just to refine the predictive accuracy but also to ensure that the insights generated by our model are equitable, generalizable, and applicable across various clinical settings.

Conclusion

In summary, Top-BERT introduces a transformative approach that deviates from the traditional pretraining-finetuning paradigm of language models, demonstrating its effective predictive performance on clinical tasks within the confines of a single-center EHR dataset, even with limited sample sizes. The model explanations using Integrated Gradients further validate the clinical applicability of our findings, highlighting its promise for enhancing diabetes care management and patient health outcomes. Moreover, our study deployed a robust framework integrating the strengths of MU EHR data lake and NRP Nautilus infrastructures—leveraging parallel high-performance computing significantly reduced model training time, boosting efficiency, and accelerating the development of our sophisticated predictive models from an

estimated 65 days to just 2 days. Furthermore, we prioritized the reproducibility of our research, making our codes accessible and clearly illustrating our methodology through informative graphics and diagrams.

Code Availability

The complete codebase for this framework is accessible through our GitHub repository: https://github.com/hikf3/task-oriented-predictive-BERT

References

1. *Center for Disease Controls and Prevention National Diabetes Statistics Report 2020: Estimates of Diabetes and Its Burden in the United States*. (2020).
2. Cicek, M., Buckley, J., Pearson-Stuttard, J. & Gregg, E. W. Characterizing Multimorbidity from Type 2 Diabetes: Insights from Clustering Approaches. *Endocrinol Metab Clin North Am* **50**, 531–558 (2021).
3. Ndjaboue, R. *et al.* Predictive models of diabetes complications: protocol for a scoping. *Syst Rev* **9**, (2020).
4. Wang, H. *et al.* Predicting hospital readmission via cost-sensitive deep learning. *IEEE/ACM transactions on computational biology and bioinformatic* **15**, 1968–1978 (2018).
5. Lysaght, T., Lim, H. Y., Xafis, V. & Ngiam, K. Y. AI-Assisted Decision-making in Healthcare: The Application of an Ethics Framework for Big Data in Health and Research. *Asian Bioeth Rev* **11**, 299–314 (2019).
6. Ahmed, Z., Mohamed, K., Zeeshan, S. & Dong, X. Q. Artificial intelligence with multi-functional machine learning platform development for better healthcare and precision medicine. *Database* **2020**, (2020).
7. Davenport, T. & Kalakota, R. The potential for artificial intelligence in healthcare. *Future Healthc J* **6**, 94 (2019).
8. Mohsen, F., Al-Absi, H., Yousri, N., El Hajj, N. & Shah, Z. A scoping review of artificial intelligence-based methods for diabetes risk prediction. *NPJ Digit Med* **6**, 197 (2023).
9. Tsao, H. Y., Chan, P. Y. & Su, E. C. Y. Predicting diabetic retinopathy and identifying interpretable biomedical features using machine learning algorithms. *BMC Bioinformatics* **19**, 30 (2018).
10. Dagliati, A. *et al.* Machine Learning Methods to Predict Diabetes Complications. *J Diabetes Sci Technol* **12**, 295–302 (2018).
11. Islam, H. & Mosa, A. S. M. A Federated Mining Approach on Predicting Diabetes-Related Complications: Demonstration Using Real-World Clinical Data. in *AMIA Annual Symposium Proceedings. Accepted. In Press.* (2021).
12. Makino, M. *et al.* Artificial intelligence predicts the progression of diabetic kidney disease using big data machine learning. *Scientific Reports 2019 9:1* **9**, 1–9 (2019).
13. Song, X. *et al.* Longitudinal Risk Prediction of Chronic Kidney Disease in Diabetic Patients using Temporal-Enhanced Gradient Boosting Machine: Retrospective Cohort Study. *JMIR Med Inform* **8**, e15510 (2020).
14. Mosa, A. S. M. *et al.* Evaluation of machine learning applications using real-world EHR data for predicting diabetes-related long-term complications. *Journal of Business Analytics* 1–11 (2021) doi:10.1080/2573234X.2021.1979901.
15. Gosak, L., Martinović, K., Lorber, M. & Stiglic, G. Artificial intelligence based prediction models for individuals at risk of multiple diabetic complications: A systematic review of the literature. *J Nurs Manag* **30**, 3765–3776 (2022).
16. Mohsen, F., Al-Absi, H., Yousri, N., El Hajj, N. & Shah, Z. A scoping review of artificial intelligence-based methods for diabetes risk prediction. *NPJ Digit Med* **6**, 197 (2023).
17. Xiao, C., Choi, E. & Sun, J. Opportunities and challenges in developing deep learning models using electronic health records data: A systematic review. *Journal of the American Medical Informatics Association* vol. 25 1419–1428 Preprint at https://doi.org/10.1093/jamia/ocy068 (2018).

18. Ayala Solares, J. R. *et al.* Deep learning for electronic health records: A comparative review of multiple deep neural architectures. *J Biomed Inform* **101**, 103337 (2020).
19. Shickel, B., Tighe, P. J., Bihorac, A. & Rashidi, P. Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis. *IEEE J Biomed Health Inform* **22**, 1589–1604 (2018).
20. Rahimian, F. *et al.* Predicting the risk of emergency admission with machine learning: Development and validation using linked electronic health records. *PLoS Med* **15**, (2018).
21. Li, Y. *et al.* BEHRT: Transformer for Electronic Health Records. *Scientific Reports 2020 10:1* **10**, 1–12 (2020).
22. Nguyen, P., Tran, T., Wickramasinghe, N. & Venkatesh, S. Deepr: A Convolutional Net for Medical Records. *IEEE J Biomed Health Inform* **21**, 22–30 (2016).
23. Choi, E., Bahadori, M., Schuetz, A., Stewart, W. F. & Sun, J. Doctor ai: Predicting clinical events via recurrent neural networks. in *Machine learning for healthcare conference* 301–318 (PMLR, 2016).
24. Pham, T., Tran, T., Phung, D. & Venkatesh, S. Deepcare: A deep dynamic memory model for predictive medicine. in *Advances in Knowledge Discovery and Data Mining: 20th Pacific-Asia Conference,PAKDD 2016, Auckland, New Zealand, April 19-22, 2016, Proceedings, Part II* 30–41 (2016).
25. Choi, E. *et al.* Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. *Adv Neural Inf Process Syst* **29**, 3504–3512 (2016).
26. Vaswani, A. *et al.* Attention Is All You Need. *Adv Neural Inf Process Syst* **2017-Decem**, 5999–6009 (2017).
27. Rasmy, L., Xiang, Y., Xie, Z., Tao, C. & Zhi, D. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ Digit Med* **4**, 1–13 (2021).
28. Rao, S. *et al.* An Explainable Transformer-Based Deep Learning Model for the Prediction of Incident Heart Failure. *IEEE J Biomed Health Inform* **26**, 3362–3372 (2022).
29. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. in *Proceedings of Human Langugage Technology: North American Chapter of the Association for Computational Linguistics (NAACL-HTL)* 4171–4186 (2019).
30. Zerka, F. *et al.* Systematic Review of Privacy-Preserving Distributed Machine Learning From Federated Databases in Health Care. *JCO Clin Cancer Inform* **4**, 184–200 (2020).
31. Curtis, L. H., Brown, J. & Platt, R. Four Health Data Networks Illustrate The Potential For A Shared National Multipurpose Big-Data Network. *Health Aff* **33**, 1178–1186 (2017).
32. Vajapeyam, S. Understanding Shannon's Entropy metric for Information. (2014).
33. Sundararajan, M., Taly, A. & Yan, Q. Axiomatic attribution for deep networks. in *International conference on machine learning.* 3319–3328 (PMLR, 2017).
34. Binder, A., Montavon, G., Lapuschkin, S., Müller, K.-R. & Samek, W. Layer-Wise Relevance Propagation for Neural Networks with Local Renormalization Layers. in *Artificial Neural Networks and Machine Learning–ICANN 2016: 25th International Conference on Artificial Neural Networks* 63–71 (Springer International Publishing, 2016). doi:10.1007/978-3-319-44781-0_8.
35. Zeiler, M. D. & Fergus, R. Visualizing and Understanding Convolutional Networks. in *Computer Vision–ECCV 2014: 13th European Conference* 818–833 (Springer International Publishing, 2014). doi:10.1007/978-3-319-10590-1_53.
36. Vaswani, A. *et al.* Attention is All you Need. *Adv Neural Inf Process Syst* **30**, (2017).
37. Furmanchuk, A. *et al.* Effect of the Affordable Care Act on diabetes care at major health centers: newly detected diabetes and diabetes medication management. *BMJ Open Diabetes Res Care* **9**, e002205 (2021).

38. Nichols, G. *et al.* Construction of a Multisite DataLink Using Electronic Health Records for the Identification, Surveillance, Prevention, and Management of Diabetes Mellitus: The SUPREME-DM Project. *Prev Chronic Dis* **9**, (2012).
39. SUPREME-DM Home. http://www.supreme-dm.org/.
40. Raebel, M. A. *et al. Mini-Sentinel Methods: Validating Type 1 And Type 2 Diabetes Mellitus In The Mini-Sentinel Distributed Database Using The Surveillance, Prevention, And Management Of Diabetes Mellitus (Supreme-DM) Datalink*. (2016).
41. De Freitas, J. K. *et al.* Phe2vec: Automated disease phenotyping based on unsupervised embeddings from electronic health records. *Patterns (N Y)* **2**, (2021).
42. Paszke, A. *et al.* Pytorch: An imperative style, high-performance deep learning library. in *Advances in Neural Information Processing Systems* 32 (2019).
43. Chen, T. & Guestrin, C. XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* **13-17-August-2016**, 785–794 (2016).
44. Ferri, C., Hernández-Orallo, J. & Flach, P. A. A coherent interpretation of AUC as a measure of aggregated classification performance. in *Proceedings of the 28th International Conference on Machine Learning (ICML-11)* 657–664 (2011).
45. Sundararajan, M., Taly, A. & Yan, Q. Axiomatic attribution for deep networks. in *International conference on machine learning.* 3319–3328 (PMLR, 2017).
46. Zhang, Y. & Yang, Q. A Survey on Multi-Task Learning. *IEEE Transactions on Knowledge and Data Engineering* vol. 34 5586–5609 Preprint at https://doi.org/10.1109/TKDE.2021.3070203 (2022).
47. Orasanu, G. & Plutzky, J. The Continuum of Diabetic Vascular Disease: From Macro- to Micro-. *J Am Coll Cardiol* **53**, S35 (2009).
48. Ohishi, M. Hypertension with diabetes mellitus: physiology and pathology. *Hypertension Research 2018 41:6* **41**, 389–393 (2018).
49. Chehade, J. M., Gladysz, M. & Mooradian, A. D. Dyslipidemia in type 2 diabetes: Prevalence, pathophysiology, and management. *Drugs* **73**, 327–339 (2013).
50. Liamis, G., Liberopoulos, E., Barkas, F. & Elisaf, M. Diabetes mellitus and electrolyte disorders. *World Journal of Clinical Cases : WJCC* **2**, 488 (2014).
51. Herrmann, M. *et al.* Serum 25-Hydroxyvitamin D: A Predictor of Macrovascular and Microvascular Complications in Patients With Type 2 Diabetes. *Diabetes Care* **38**, 521–528 (2015).
52. Williams, A. *et al.* Pathophysiology of Red Blood Cell Dysfunction in Diabetes and Its Complications. *Pathophysiology 2023, Vol. 30, Pages 327-345* **30**, 327–345 (2023).
53. Association, A. D. 11. Microvascular Complications and Foot Care: Standards of Medical Care in Diabetes−2020. *Diabetes Care* **43**, S135–S151 (2020).
54. Metzger, M., Abdel-Rahman, E. M., Boykin, H. & Song, M. K. A Narrative Review of Management Strategies for Common Symptoms in Advanced CKD. *Kidney Int Rep* **6**, 894–904 (2021).
55. Ravaut, M. *et al.* Predicting adverse outcomes due to diabetes complications with machine learning using administrative health data. *npj Digital Medicine 2021 4:1* **4**, 1–12 (2021).
56. Kim, E., Pieczkiewicz, D. S., Castro, M. R., Caraballo, P. J. & Simon, G. J. Multi-Task Learning to Identify Outcome-Specific Risk Factors that Distinguish Individual Micro and Macrovascular Complications of Type 2 Diabetes. *AMIA Summits on Translational Science Proceedings* **2018**, 122 (2018).

Acknowledgements

The data set used for the analyses described were obtained from PCORnet Common Data Model at the University of Missouri, which is funded by the Patient Centered Outcomes Research Institute PCORnet (RI-MISSOURI-01-PS1) Clincal Research Network, the Greater Plains Collaborative. The NSF Nautilus Kubernetes HyperCluster is supported by the University of Missouri Greater Plains Regional CyberTeam (NSF Award #1925681).

Author contributions

HI: conceptualization, methodology, data analysis, clinical interpretations, and writing.
GB: clinical interpretations, validation, review, and editing.
RP: clinical interpretations, validation, review, and editing.
PR: methodology, validation, review, and editing.
RW: clinical interpretations, validation, review, and editing.
XS: supervision, methodology, clinical interpretations, validation, review, and editing.

Competing interests

The authors declare no competing interests.

Supplementary Document

Comparing BERT-based Model Performances: Table S1 shows the micro-averaged performance metrics for two finetuning frameworks and Top-BERT embedding configurations. Metrics were computed for epoch 200 for the two finetuning frameworks after pretraining for 500 epochs. For the Top-BERT models all metrics were computed at epoch 350. We also compared the AUROC of prolonged hospital stay in the Top-BERT models, as shown in Table S1.

	Finetuning A							
	Input	**Embedding**	micro-auroc	micro-recall	micro-precision	micro-f1	epoch	
1	DX	age	0.6789	0.7992	0.7914	0.7951	200	
2	DX	time diff	0.6957	0.7828	0.8010	0.7911	200	
3	DX	visit	0.6709	0.7998	0.8005	0.8002	200	
4	DX	pos+seg	0.6745	0.7917	0.7907	0.7912	200	
5	DX	age+time diff	0.6725	0.7978	0.7906	0.6725	200	
6	DX	pos+seg+age	0.6440	0.7900	0.7901	0.6991	200	
7	DX	pos+seg+time diff	0.6858	0.8113	0.7998	0.8052	200	
8	DX	visit+age	0.6591	0.7961	0.7847	0.7901	200	
9	DX	visit+age+time diff	0.6568	0.8179	0.7954	0.8051	200	
10	DX	visit+time diff	0.6786	0.8023	0.7960	0.7990	200	
11	DX	pos+seg+visit	0.6754	0.7941	0.7904	0.7922	200	
12	DX	pos+seg+visit+age	0.6440	0.8175	0.7853	0.7985	200	
13	DX	pos+seg+visit+time diff	0.6491	0.8103	0.7914	0.7998	200	
14	DX	pos+seg+age+time diff	0.6501	0.8241	0.7945	0.8063	200	
15	DX	pos+seg+visit+age+time diff	0.6594	0.8197	0.7907	0.8025	200	
	Finetuning B							
	Input	**Embedding**	micro-auroc	micro-recall	micro-precision	micro-f1	epoch	
1	DX	age	0.6859	0.8053	0.7904	0.7973	200	
2	DX	time diff	0.7035	0.7948	0.8014	0.7980	200	
3	DX	visit	0.7086	0.8071	0.7995	0.8009	200	
4	DX	pos+seg	0.6939	0.8032	0.7802	0.7952	200	
5	DX	age+time diff	0.6641	0.8178	0.7888	0.8008	200	
6	DX	pos+seg+age	0.6652	0.8125	0.7838	0.7957	200	
7	DX	pos+seg+time diff	0.6903	0.8143	0.7906	0.8008	200	
8	DX	visit+age	0.6725	0.8007	0.7901	0.7952	200	
9	DX	visit+age+time diff	0.6869	0.8191	0.7950	0.8051	200	
10	DX	visit+time diff	0.6890	0.804	0.7978	0.8008	200	
11	DX	pos+seg+visit	0.6876	0.8080	0.7917	0.7991	200	
12	DX	pos+seg+visit+age	0.6705	0.830	0.7886	0.8033	200	
13	DX	pos+seg+visit+time diff	0.6898	0.8242	0.7947	0.8061	200	
14	DX	pos+seg+age+time diff	0.7057	0.8257	0.7981	0.8090	200	
15	DX	pos+seg+visit+age+time diff	0.6713	0.8065	0.7830	0.7933	200	
	Top-BERT							
	Input	**Embedding**	micro-auroc	micro-recall	micro-precision	micro-f1	epoch	Prolonged LOS auroc
1	DX	age	0.6759	0.8476	0.8709	0.8585	350	0.7331
2	DX	time diff	0.6947	0.7607	0.8802	0.8079	350	0.8011
3	DX	visit	0.7105	0.7963	0.8825	0.8815	350	0.7958

4	DX	pos+seg	0.6933	0.8155	0.8766	0.8419	350	0.7602
5	DX	age+time diff	0.6850	0.8489	0.8716	0.8596	350	0.7413
6	DX	pos+seg+age	0.6753	0.8761	0.8690	0.8725	350	0.7449
7	DX	pos+seg+time diff	0.6976	0.8591	0.8756	0.8669	350	0.7468
8	DX	visit+age	0.6795	0.8589	0.8719	0.8652	350	0.7577
9	DX	visit+age+time diff	0.6796	0.8725	0.8690	0.8707	350	0.7619
10	DX	visit+time diff	0.7125	0.7839	0.8815	0.8233	350	0.7699
11	DX	pos+seg+visit	0.6960	0.8317	0.8757	0.8513	350	0.7585
12	DX	pos+seg+visit+age	0.6751	0.8699	0.8682	0.8690	350	0.7470
13	DX	pos+seg+visit+time diff	0.693	0.8619	0.8749	0.8681	350	0.7399
14	DX	pos+seg+age+time diff	0.6719	0.8669	0.8693	0.8681	350	0.7634
15	DX	pos+seg+visit+age+time diff	0.6766	0.8641	0.8694	0.8667	350	0.7487

Table S 1. Comparing the micro-averaged metrics for each embedding configurations for the BERT-based model frameworks. Additionally, the AUROC of prolonged hospital stay is provided for the Top-BERT models.

Comparing Task-Specific Model Performances:

To gain insights on the differences and similarities of Top-BERT and XGBoost in handling class imbalance, we compared the model performance for each classification task: CKD, MACE, NEUR, and RET using precision, recall, F1-score, Mathew's correlation coefficient (MCC) and confusion matrix computed at varying threshold values (10%, 30%, 50%, and 80%). Our test data consisted of 7.25% of CKD labels, 11.6% of MACE, 10.9% of NEUR, and 2.7% of RET. Table S2 compares the metrics for each classification task for five Top-BERT embedding configurations (having comparable mAUCs) and XGBoost. Top-BERT models displayed more balanced metrics across the thresholds with significantly higher true positive predictions, leading to higher recall values than XGBoost. However, XGBoost showed lower false positive rates, contributing to higher precision, particularly at higher thresholds than Top-BERT. For instance, XGBoost achieved a precision of 69.5% for predicting MACE at 50% threshold compared to 26.4% in Top-BERT (input+time_diff). False positives of MACE predicted by XGBoost were significantly lower than Top-BERT models (0.28% for XGBoost vs. 9.32% for Top-BERT input+time_diff) — which contributed to the significant increase in precision for XGBoost. However, XGBoost identified 41 true positives out of the 849 positive labels, which resulted in a recall value of 4.8% compared to 414 true positives (recall of 48.8%) for Top-BERT (input+time_diff). Additionally, to predict RET, which had the lowest fraction of true positives, true positive rates for XGBoost were zero in most cases, whereas input+time_diff identified 21.5% true positives at the 50% threshold. Thus, we observed that higher precision values and lower recall values obtained by XGBoost were consequences of conservatively predicting fewer number of positive cases (both true positives and false positives) in all classification tasks. On the other hand, Top-BERT showed better measure of separability, as indicated by higher mAUC values, and higher efficiency in predicting the positive classes compared to XGBoost.

Chronic Kidney Disease (CKD)									**Major Adverse Cardiac Events (MACE)**								
Model			**input+ visit+ time_diff**	**input+ visit**	**input+ visit+ pos+ seg**	**input+ time_diff**	**input+ pos+ seg+ time_diff**	**XGBoost**	**Model**			**input+ visit+ time_diff**	**input+ visit**	**input+ visit+ pos+ seg**	**input+ time_diff**	**input+ pos+ seg+ time_diff**	**XGBoost**
Chronic Kidney Disease (CKD)	**Precision**	**10%**	0.1295	0.1170	0.1284	0.1081	0.1520	0.1965	**Major Adverse Cardiac Events (MACE)**	**Precision**	**10%**	0.1792	0.1691	0.1798	0.1620	0.2331	0.2226
		30%	0.1441	0.1266	0.1391	0.1122	0.1605	0.3462			**30%**	0.1970	0.1859	0.2009	0.1791	0.2470	0.4501
		50%	0.1537	0.1314	0.1458	0.1205	0.1655	0.7273			**50%**	0.2116	0.2111	0.2147	0.1893	0.2641	0.6949
		80%	0.1678	0.1575	0.1696	0.1377	0.1788	0.0000			**80%**	0.2400	0.2595	0.2459	0.2354	0.2922	0.6667
	Recall	**10%**	0.4415	0.5906	0.4113	0.5792	0.2472	0.4943		**Recall**	**10%**	0.6125	0.6737	0.5112	0.6643	0.3604	0.7173

		30%	0.3755	0.4792	0.3170	0.4943	0.2057	0.0509
		50%	0.3208	0.3943	0.2566	0.4377	0.1774	0.0151
		80%	0.2264	0.2642	0.1660	0.3019	0.1302	0.0000
	F-1	**10%**	0.2003	0.1953	0.1957	0.1822	0.1882	0.2813
		30%	0.2083	0.2003	0.1933	0.1829	0.1803	0.0888
		50%	0.2078	0.1972	0.1859	0.1889	0.1712	0.0296
		80%	0.1928	0.1973	0.1678	0.1891	0.1507	0.0000
	MCC	**10%**	0.1259	0.1303	0.1185	0.1093	0.1120	0.2259
		30%	0.1332	0.1283	0.1142	0.1048	0.1086	0.1096
		50%	0.1322	0.1198	0.1080	0.1105	0.1041	0.0980
		80%	0.1210	0.1219	0.1034	0.1092	0.0967	0.0000
	True Positive	**10%**	234	313	218	307	131	262
		30%	199	254	168	262	109	27
		50%	170	209	136	232	94	8
		80%	120	140	88	160	69	0
	False Positive	**10%**	1573	2362	1480	2533	731	1071
		30%	1182	1752	1040	2073	570	51
		50%	936	1381	797	1694	474	3
		80%	595	749	431	1002	317	0
	True Negative	**10%**	5204	4415	5297	4244	6046	5706
		30%	5595	5025	5737	4704	6207	6726
		50%	5841	5396	5980	5083	6303	6774
		80%	6182	6028	6346	5775	6460	6777
	False Negative	**10%**	296	217	312	223	399	268
		30%	331	276	362	268	421	503
		50%	360	321	394	298	436	522
		80%	410	390	442	370	461	530

Neuropathy (NEUR)

Model	**input+ visit+ time_diff**	**input+ visit**	**input+ visit+ pos+ seg**	**input+ time_diff**	**input+ pos+ seg+ time_diff**	**XGBoost**

		30%	0.5253	0.5453	0.4158	0.5783	0.2933	0.1861
		50%	0.4653	0.4629	0.3569	0.4876	0.2544	0.0483
		80%	0.3475	0.2733	0.2627	0.3145	0.2038	0.0024
	F-1	**10%**	0.2773	0.2704	0.266	0.2604	0.2831	0.3397
		30%	0.2865	0.2772	0.2709	0.2735	0.2682	0.2633
		50%	0.2909	0.2899	0.2681	0.2727	0.2591	0.0903
		80%	0.2839	0.2662	0.254	0.2693	0.2401	0.0047
	MCC	**10%**	0.1596	0.1534	0.1394	0.1363	0.1707	0.2569
		30%	0.1690	0.1564	0.1488	0.1522	0.1633	0.2341
		50%	0.1744	0.1731	0.1504	0.1491	0.1638	0.1629
		80%	0.1738	0.1670	0.1523	0.1595	0.1631	0.0348
	True Positive	**10%**	520	572	434	564	306	609
		30%	446	463	353	491	249	158
		50%	395	393	303	414	216	41
		80%	295	232	223	267	173	2
	False Positive	**10%**	2382	2810	1980	2918	1007	2127
		30%	1818	2028	1404	2250	759	193
		50%	1472	1469	1108	1773	602	18
		80%	934	662	684	867	419	1
	True Negative	**10%**	4076	3648	4478	3540	5451	4331
		30%	4640	4430	5054	4208	5699	6265
		50%	4986	4989	5350	4685	5856	6440
		80%	5524	5796	5774	5591	6039	6457
	False Negative	**10%**	329	277	415	285	543	240
		30%	403	386	496	358	600	691
		50%	454	456	546	435	633	808
		80%	554	617	626	582	676	847

Retinopathy (RET)

Model	**input+ visit+ time_diff**	**input+ visit**	**input+ visit+ pos+ seg**	**input+ time_diff**	**input+ pos+ seg+ time_diff**	**XGBoost**

Neuropathy (NEUR)	**Precision**	**10%**	0.17192	0.1565	0.1648	0.1508	0.1971	0.2256	**Retinopathy (RET)**	**Precision**	**10%**	0.0863	0.0894	0.0933	0.0730	0.1058	0.1552
		30%	0.18289	0.1839	0.1858	0.1655	0.2144	0.4778			**30%**	0.0909	0.0959	0.0957	0.0786	0.1095	0.0000
		50%	0.19700	0.2044	0.2059	0.1803	0.2292	0.6939			**50%**	0.0993	0.1005	0.0935	0.0883	0.1221	0.0000
		80%	0.21867	0.2505	0.2468	0.2075	0.2521	1.0000			**80%**	0.1096	0.1020	0.0829	0.0978	0.1151	0.0000
	Recall	**10%**	0.60453	0.6927	0.5567	0.7028	0.4572	0.6675		**Recall**	**10%**	0.18	0.235	0.195	0.265	0.145	0.215
		30%	0.48992	0.5945	0.4723	0.5957	0.3741	0.2166			**30%**	0.15	0.21	0.155	0.22	0.11	0
		50%	0.43073	0.5126	0.4207	0.5076	0.3325	0.0856			**50%**	0.145	0.19	0.13	0.215	0.105	0
		80%	0.31864	0.3476	0.3174	0.3489	0.2670	0.0063			**80%**	0.125	0.15	0.085	0.175	0.08	0
	F-1	**10%**	0.26771	0.2553	0.2543	0.2483	0.2754	0.3373		**F-1**	**10%**	0.1167	0.1295	0.1262	0.1145	0.1224	0.1803
		30%	0.26635	0.2809	0.2667	0.2590	0.2726	0.2981			**30%**	0.1132	0.1317	0.1183	0.1158	0.1097	0
		50%	0.27036	0.2923	0.2765	0.2661	0.2713	0.1525			**50%**	0.1179	0.1315	0.1088	0.1252	0.1129	0
		80%	0.25935	0.2911	0.2777	0.2602	0.2593	0.0125			**80%**	0.1168	0.1215	0.0840	0.1254	0.0944	0
	MCC	**10%**	0.15984	0.1480	0.1374	0.1371	0.1649	0.2587		**MCC**	**10%**	0.0889	0.1058	0.0995	0.0929	0.0949	0.1556
		30%	0.15283	0.1778	0.1532	0.1464	0.1644	0.2700			**30%**	0.0847	0.1060	0.0902	0.0904	0.0846	-0.0048
		50%	0.15846	0.1883	0.1669	0.1528	0.1675	0.2192			**50%**	0.0900	0.1047	0.0806	0.0998	0.0901	0
		80%	0.15331	0.1920	0.1789	0.1501	0.1662	0.0749			**80%**	0.0905	0.0937	0.0578	0.0979	0.0749	0
	True Positive	**10%**	480	550	442	558	363	530		**True Positive**	**10%**	36	47	39	53	29	43
		30%	389	472	375	473	297	172			**30%**	30	42	31	44	22	0
		50%	342	407	334	403	264	68			**50%**	29	38	26	43	21	0
		80%	253	276	252	277	212	5			**80%**	25	30	17	35	16	0
	False Positive	**10%**	2312	2964	2240	3143	1479	1819		**False Positive**	**10%**	381	479	379	673	245	234
		30%	1738	2095	1643	2385	1088	188			**30%**	300	396	293	516	179	6
		50%	1394	1584	1288	1832	888	30			**50%**	263	340	252	444	151	0
		80%	904	826	769	1058	629	0			**80%**	203	264	188	323	123	0
	True Negative	**10%**	4201	3549	4273	3370	5034	4694		**True Negative**	**10%**	6726	6628	6728	6434	6862	6873
		30%	4775	4418	4870	4128	5425	6325			**30%**	6807	6711	6814	6591	6928	7101
		50%	5119	4929	5225	4681	5625	6483			**50%**	6844	6767	6855	6663	6956	7107
		80%	5609	5687	5744	5455	5884	6513			**80%**	6904	6843	6919	6784	6984	7107
	False Negative	**10%**	314	244	352	236	431	264		**False Negative**	**10%**	164	153	161	147	171	157
		30%	405	322	419	321	497	622			**30%**	170	158	169	156	178	200
		50%	452	387	460	391	530	726			**50%**	171	162	174	157	179	200
		80%	541	518	542	517	582	789			**80%**	175	170	183	165	184	200

Table S 2. Comparing the task-specific model performances for five Top-BERT embedding configurations (having comparable micro-averaged AUROCs) and XGBoost for each classification task: CKD, MACE, NEUR, and RET using precision, recall, F1-score, Mathew's correlation coefficient (MCC) and confusion matrix computed at varying threshold values (10%, 30%, 50%, and 80%). The red highlighted scores show the highest value in each metric for each threshold value.