2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) | 979-8-3503-3748-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/BIBM58861.2023.10385529

Sequential Representation of Sparse Heterogeneous Data for Diabetes Risk Prediction

Rochana Chaturvedi[1], Mudassir Rashid[2], Brian T. Layden[1], Andrew Boyd[1], Ali Cinar[2], Barbara Di Eugenio[1]
[1]University of Illinois Chicago, USA [2]Illinois Institute of Technology, Chicago, USA
[1]{rchatu2, blayde1, boyda, bdieugen}@uic.edu, [2]{mrashid3, cinar}@iit.edu

***Abstract*—Type 2 diabetes (T2D) is a major public health problem, and opportunistic screening to detect T2D at an early stage can help initiate interventions that delay or prevent the disease and its complications. In this study, we use electronic health records (EHR) and concepts extracted from clinical notes to predict future T2D risk. Our deep neural network-based model captures the temporal sequence of patient visits. We use explainable AI algorithms to assess the model decisions and observe alignment with the domain knowledge of clinical experts.**

***Index Terms*—Machine Learning, Natural Language Processing, Disease Prediction, Diabetes**

I. Introduction

Type 2 diabetes (T2D) is a highly prevalent chronic condition, affecting approximately 35 million people in the United States, 23% of whom are undiagnosed [1]. Early detection of T2D, coupled with opportunistic screening, can identify patients who are likely to benefit from lifestyle interventions (diet and exercise) and medications and prevent serious complications. This is even more important for the underprivileged populations served by UI Health, the hospital at the institution we are affiliated with—often, these patients do not access the health system on a regular basis, but rather, occasionally visit a doctor or the emergency room when a health crisis occurs.

Traditionally, physicians conduct manual reviews of clinical notes to identify patients at risk of developing T2D. However such an approach is not readily scalable. Predictive models trained using data from electronic health records (EHR) (such as diagnostic codes, patient demographics, vital signs, laboratory tests, and prescribed medications) can increase patient access to diabetes screening [2]. An important and abundant source of information in EHRs is the unstructured data comprising notes written by healthcare providers—descriptions elucidating laboratory test results, inpatient discharge summaries, etc. Leveraging the wealth of information contained in these notes requires the use of Natural Language Processing (NLP).

Additionally, EHRs contain longitudinal information in the form of multiple patient visits over time. This time-series data contains hidden patterns and latent time-varying features on the progression of symptoms and other complications that can be exploited to predict the future risk of disease [3], [4]. However, there has been limited work on modeling the information from clinical notes as a time series owing to challenges such as irregularly distributed events, incongruent and fragmented notes across visits and healthcare providers, and data heterogeneity and sparsity. In this work, we propose a novel neural architecture to model the temporal sequence of clinical notes across patient visits as an irregularly sampled time series and predict the future risk of T2D. We effectively integrate the distinct modalities of unstructured notes across different patient visits and structured data in the proposed framework. We explicitly handle the temporal information in the sequence of clinic visits using recurrent neural networks. We analyze the concepts and features that are most predictive and those that are least predictive of the future risk of T2D.

II. Data Description

We begin with a large dataset comprising adults (age $\geq$18 years) who have been treated at UI Health from January 1, 2010, to July 31, 2021.[1] The data contains UI Health patients with and without a diagnosis of T2D. The exclusion criteria include diagnosis of type 1 diabetes or gestational diabetes. The data comprises clinical notes and structured variables such as demographic attributes, lab values, diagnosis dates, etc.

A. Sampling and Preprocessing

After consulting the clinical experts, we select the following most informative note types ‘Family Medicine Note’, ‘Transplant Surgery Note’, ‘History and Physical Note’, ‘Endocrinology Note’, ‘Emergency Department Note’, ‘Diabetes Education’, ‘Child Psych Discharge Summary’, ‘Exercise/Stress Procedure’. We drop the duplicate entries and undersample the non-diabetic patients by 25% so they are around twice the number of diabetic patients. Since we are only interested in a patient’s first diagnosis, we drop all visits including and after the first diagnosis. The final dataset consists of 10621 unique patients, 826 of whom are diabetic. Despite an initial undersampling of the non-diabetic group, we end up with further unbalanced data owing to many of the T2D patients being pre-diagnosed relative to the data collection period.

B. Concept Extraction from free text

We follow a normalized-concept-based approach wherein we extract clinical concepts from text such as medications, disease, symptoms, procedures, and anatomical sites, and map them to unique identifiers (CUIs) in the Unified Medical Language System (UMLS) [5] metathesaurus using NLP-based tool cTAKES [6]. Fig. 1 shows an example of concepts extracted from our data. Minus sign indicates negation of a concept.

[1] The data is confidential due to protected health information.

979-8-3503-3748-8/23/$31.00 ©2023 IEEE

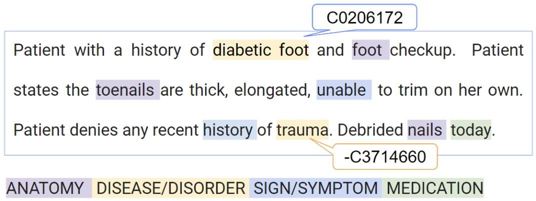

C0206172
Patient with a history of diabetic foot and foot checkup. Patient states the toenails are thick, elongated, unable to trim on her own. Patient denies any recent history of trauma. Debrided nails today.
-C3714660
ANATOMY
DISEASE/DISORDER
SIGN/SYMPTOM
MEDICATION

Fig. 1: Concepts CUIs extracted by cTAKES

C. Diabetes Progression Time-Series

In addition to the concepts extracted from clinical notes, we also experiment with a subset of structured variables. These include sensitive attributes such as gender—Female (F) or Male (M), race—Black (B), White (W), or No Information (NI), ethnicity—Hispanic (Y) or non-Hispanic (N), and age group—age 18-45 as Adults (A), 46-65 as Middle Aged (M), and 65 and above as Seniors (S) shown in Fig. 2.[2] We also use the HbA1c test values before the diagnosis, shown in Fig. 3. This is a weighted average of blood glucose over the past 3 months which facilitates the diagnosis of T2D at value $\geq 6.5\%$ and prediabetes if it is between 5.7 and 6.4. This is a sparse data source, missing from around 56.8% observations. There are several patients in the non-diabetic group with HbA1c $\geq 6.5\%$ without a clinical diagnosis of T2D, indicative of potential missed diagnoses.

For each patient, we combine all their visits into a single observation and sort them from first to last before diagnosis. The number of visits varies from 2–234 with a mean of 4.7. We use up to the last ten visits (91st percentile of total visits) before diagnosis and pad the shorter sequences using a special token. We split the data into training-validation-test sets stratified by the T2D diagnosis, in the ratio 80:10:10.

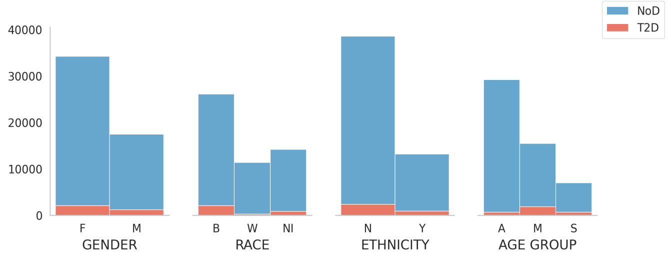


Fig. 2: Demographics of T2D and non-Diabetic (NoD) patients.

III. Methodology

A. Baselines

We use Logistic Regression (LR) with 10-fold cross-validation (CV) over the combined training and validation splits. For the first baseline, we use the last HbA1c value before a diagnosis, imputing missing values with the mean. As second baseline, we use the sensitive attributes—race, ethnicity, gender, and age group. And as third, a combination of both.

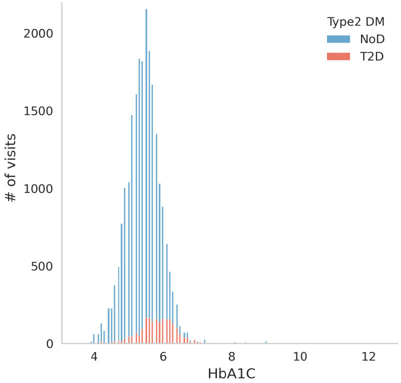


Fig. 3: Last HbA1c value before diagnosis in our data.

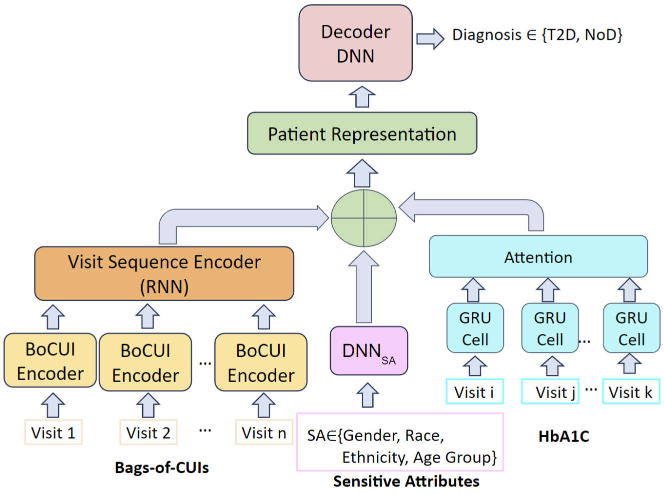


Fig. 4: Diabetes Progression Time-Series Model Architecture

B. Concept-based Models

a) LR with TF-IDF We flatten the visit sequence by combining all the unique concepts from the last ten visits into a single bag-of-CUIs. We train an LR model with L2 regularization and 10-fold cross-validation using Term Frequency-Inverse Document Frequency (TF-IDF) vectorization with 1–2 n-grams.

b) Sequential Models: We propose a neural architecture comprising two encoders—a visit level encoder for each bag-of-CUIs followed by a patient level encoder for the sequence of visits. Each CUI is mapped to a vector embedding—random initialization vs. pre-trained cui2vec embeddings [7] and fine-tuned during model training.[3]

i) CNN-LSTM: As the first encoder pair, we use a Convolutional Neural Network (CNN) to encode the bag-of-CUIs in each visit followed by a Long Short Term Memory (LSTM) network to encode the sequence of visits.[4]

ii) Hierarchical Attention Network (HAN): We combine stacks of recurrent neural networks (RNN) and attention layers

[2] The distinction in racial categories and the Hispanic/Latino ethnicity follow from the US census.

[3] We were able to map 17360 out of 24701 CUIs in the training set to the pre-trained vectors and randomly initialized those that couldn't be mapped.

[4] We also experimented with mean CUI embeddings (MCE), MCE with attention, deep averaging network (DAN), and DAN with attention that do not perform well.

using HAN [8] at two granularities—visit level and patient level. We experiment with unidirectional and bidirectional LSTM and Gated Recurrent Units (GRU) as choices of RNN.

C. Multi-modal Models

We also combine text representation with structured data. For sensitive attributes, we transform each attribute type with a fully connected neural network layer specific to that attribute to get attribute representation. To model the HbA1c time series, we use a sub-network consisting of GRU with attention mechanism. The final representations of text and structured data are concatenated together and processed via a fully connected layer with sigmoid activation to get the probability of T2D diagnosis. The model architecture is represented in Fig. 4.

The models are trained for 20 epochs using a batch size of 32 on RTX 2080 Ti GPU using Keras API. We reduce the learning rate if the validation loss doesn't improve for a few epochs. We use Adam optimizer to minimize binary cross-entropy loss and class weights to handle imbalance. The kernel weights are randomly initialized. Masking is used to model missing values in HbA1c time series, batch normalization is used for scaling, and dropout is used to prevent overfitting. We use Bayesian search and manual tuning to select the hyperparameters and save the model with the least validation loss.

IV. Results and Discussion

We report the evaluation results in Table I. We use Precision (P) and Recall (R) for each of the two classes and the macro-average of their harmonic mean (macro-F1) for comparisons. Even with a majority baseline classifying all patients as non-diabetic, we achieve a high macro-average F_1 at 47.97%. Among the baselines that use only the structured or semi-structured time series inputs, we find that HbA1c is quite important. However, the last HbA1c before diagnosis is not predictive enough which can be attributed to it being sparse and noisy. The performance of the model with GRU and attention mechanism gets a notable boost owing to more balanced predictions for both classes. The models that use the sensitive attributes have poor performance. Moreover, they also degrade the performance of the last HbA1c-based model when combined with it.

In the second panel of Table I, describing CUI-based models, there is a performance improvement starting with LR. Neural models with pre-trained cui2vec embeddings provide a sizable improvement over this. Although, random initialization of CUIs leads to a performance drop. The CNN-LSTM model with cui2vec has the best performance with a macro average F1 at 74.15 due to more balanced scores for both classes.

In the third panel of Table I, we show results for bimodal models. Again, the sensitive attributes reduce the performance of the models in comparison to the respective variants in the second panel. The HbA1c values help improve the precision for the T2D class for the CNN-LSTM variant and that of the majority non diabetes class for HAN variant. However, overall this sparse input reduces the performance compared to CUI-only models.

TABLE I: Results on the held-out test set. SA refers to sensitive attributes. HbA1C$_{ts}$ refers to the time series of HbA1c values

	Model	**Text**	**Macro**	**T2D**		**NoD**	
		Encoding	**F1**	**P**	**R**	**P**	**R**
Baseline	LR (HbA1C$_{last}$)	NA	60.1	22.1	45.8	95.0	86.3
	LR (SA)	NA	54.1	16.8	72.3	96.7	69.6
	LR (HbA1C$_{last}$+SA)	NA	58.4	19.8	60.2	95.9	79.4
	RNN+Attention (HbA1C$_{ts}$)	NA	65.3	35.2	37.4	94.7	94.2
CUI	LR	tf-idf	68.2	50.9	33.7	94.5	97.2
	HAN	Random	64.3	28.1	51.8	95.6	88.8
	CNN-LSTM	Random	66.4	29.3	**73.5**	97.4	85.0
	CNN-LSTM	Cui2Vec	**74.2**	46.0	62.7	96.7	93.8
	HAN	Cui2Vec	70.7	38.8	60.2	96.5	92.0
Bimodal	HAN (CUI+SA)	Cui2Vec	69.5	43.9	43.4	95.2	95.3
	CNN-LSTM (CUI+HbA1C$_{ts}$)	Cui2Vec	69.7	**59.6**	33.7	94.6	**98.1**
	HAN (CUI+HbA1C$_{ts}$)	Cui2Vec	68.2	32.4	69.9	**97.2**	87.7
	HAN (CUI+HbA1C$_{ts}$+SA)	Cui2Vec	65.8	29.0	66.3	96.8	86.2
	Support		1063	83		980	

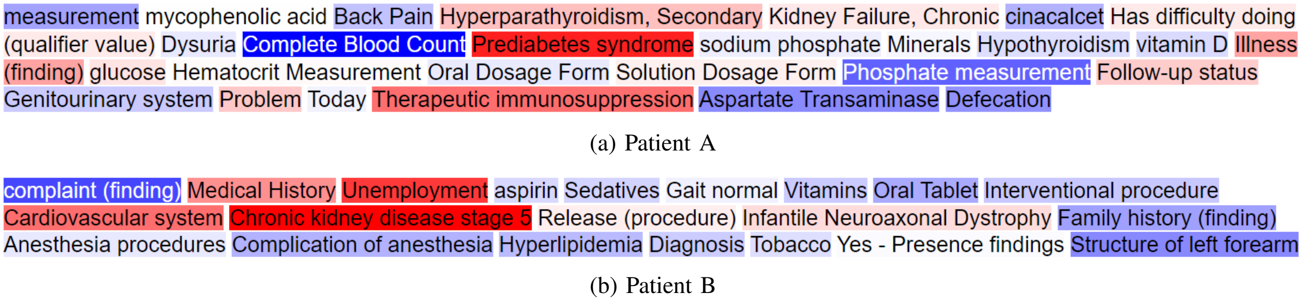
TABLE II: Top 5 features for positive (T2D) and negative (NoD) classes in the LR Model.

Positive Features		Negative Features	
CUI	Concept Name	CUI	Concept Name
C0025598	Metformin	C0747752	Polysubstance abuse
C0017642	Glipizide	C0006684	Calcium Channel Blockers
C0857112	Bilateral glaucoma	C0162703	Pain Threshold
C0591573	Glucophage	C0392557	Nuclear cataract
C0584640	Tibial plateau structure	C0178316	Fracture of upper limb

In the high-stakes setting of the public health domain, it is imperative to understand why a model predicts a certain class for a given input. Beginning with the white box LR classifier, it is straightforward to infer the features associated with each class using the coefficients corresponding to each feature (CUI). We compute the idf weighted scores from the coefficients associated with each CUI. Table II shows the top five CUIs most predictive of a positive (T2D) and negative (No Diabetes) classification respectively in decreasing order of their predictive power. Some of the top positively correlated features are antidiabetes drugs such as 'Metformin', 'Glipizide', and 'Glucophage'. Note that these are administered before the formal diagnosis. There is no known association between some other concepts in this table with T2D. For instance, undiagnosed T2D might lead to 'Bilateral glaucoma'. However, the latter is not a known cause of the prior.

To open up the black box of our best-performing CNN-LSTM classifier, we borrow the layerwise relevance propagation (LRP) [9] technique from explainable AI. LRP is a powerful technique that can uncover for each input feature (in our case, a CUI) if it has a positive or a negative contribution towards a particular prediction. These contributions are called relevance scores. LRP computes these scores by decomposing the prediction scores from the model output back to the input source by propagating it layer-by-layer, following a layerwise conservation principle. This means the relevance scores stay the same across layers. We use the LRP implementation from DeepExplain package [10][5]. We show two example heatmaps obtained with this technique in Fig. 5. The CUIs are translated

[5] https://github.com/marcoancona/DeepExplain



measurement mycophenolic acid Back Pain Hyperparathyroidism, Secondary Kidney Failure, Chronic cinacalcet Has difficulty doing (qualifier value) Dysuria Complete Blood Count Prediabetes syndrome sodium phosphate Minerals Hypothyroidism vitamin D Illness (finding) glucose Hematocrit Measurement Oral Dosage Form Solution Dosage Form Phosphate measurement Follow-up status Genitourinary system Problem Today Therapeutic immunosuppression Aspartate Transaminase Defecation

(a) Patient A

complaint (finding) Medical History Unemployment aspirin Sedatives Gait normal Vitamins Oral Tablet Interventional procedure Cardiovascular system Chronic kidney disease stage 5 Release (procedure) Infantile Neuroaxonal Dystrophy Family history (finding) Anesthesia procedures Complication of anesthesia Hyperlipidemia Diagnosis Tobacco Yes - Presence findings Structure of left forearm

(b) Patient B

Fig. 5: Normalized LRP attribution score heatmaps for CNN-LSTM from snippets of concepts for two different patients' visits.

to the corresponding concept names using UMLS rest API for readability. The image only depicts the model attributions over a small fragment from a single visit from a patient in the positive (T2D) class. The CUIs highlighted in red depict highly positive relevance scores concerning the T2D class while the ones in blue depict highly negative relevance scores. We note from Fig. 5 (a) that the concept 'Prediabetes syndrome' contributes highly towards a positive prediction. 'Therapeutic immunosuppression' also has a high positive contribution, as do 'illness' and 'glucose'. On the other hand, generic concepts such as complete blood count have highly negative relevance towards diabetes prediction. In the case of patient B, we find a different set of concepts namely 'Chronic kidney disease stage 5' as having highly positive contributions. Interestingly, our model also finds 'unemployment' to be a positive contributor. Although we cannot infer from this information alone whether and how unemployment may cause diabetes or whether symptoms leading to diabetes may cause unemployment, employment and job security are identified by the World Health Organization among the important social determinants of health.

V. Conclusion

We predict the future risk of T2D by modeling concepts extracted from the temporal sequence of patient visits using neural models. A challenge for modeling our dataset is a high class imbalance, which is reflective of the real-world setting. Despite this, we find sizable performance improvements even for the sparse T2D class by leveraging the concepts from clinical notes using sequential representation. The models' performances drop when we combine the concepts with structured attributes, one of which (HbA1c) is noisy and sparse. In the future, we will experiment with additional data sources such as glucose sensor variables, insulin pump values, patient vitals, and medications when available.

We also uncover several mistakes made by the concept extraction and normalization pipeline of cTAKES. For instance, the word "Plan" is mapped to disease/disorder type while "today" is mapped to medication. In many examples, "DM2" is mapped to the CUI for Dystrophia Myotonica Type 2 when it refers to Diabetes Mellitus Type 2. In the future, we would like to integrate other text representations, such as word embeddings and large pre-trained clinical language model embeddings.

Finally, we investigate the patterns driving model decisions for both LR and top-performing neural models using explainable AI techniques. We illustrate that the highly predictive features align with the medical knowledge, fostering trust in the models. The heatmap visualizations for a classifier can also be a useful tool for clinicians to help focus their attention on relevant concepts while reviewing a patient's medical history.

Acknowledgments

Thanks to the reviewers for their suggestions, to Sugat Chaturvedi, Ayomide Owoyemi, Kai Shu, and Canyu Chen for their help; to a DPI Science team award, and to NSF grant IIS-2312862 for financial support.

References

[1] "Centers for Disease Control and Prevention. National Diabetes Statistics Report, howpublished = https://www.cdc.gov/diabetes/data/statistics-report/index.html, note = Accessed: 2023-08-14."

[2] L. Kopitar, P. Kocbek, L. Cilar, A. Sheikh, and G. Stiglic, "Early detection of type 2 diabetes mellitus using machine learning-based prediction models," *Scientific reports*, vol. 10, no. 1, p. 11981, 2020.

[3] Z. C. Lipton, D. C. Kale, C. Elkan, and R. C. Wetzel, "Learning to diagnose with LSTM recurrent neural networks," in *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2016. [Online]. Available: http://arxiv.org/abs/1511.03677

[4] H. Wang, Y. Li, S. A. Khan, and Y. Luo, "Prediction of breast cancer distant recurrence using natural language processing and knowledge-guided convolutional neural network," *Artificial intelligence in medicine*, vol. 110, p. 101977, 2020.

[5] O. Bodenreider, "The unified medical language system (umls): integrating biomedical terminology," *Nucleic acids research*, vol. 32, no. suppl_1, pp. D267–D270, 2004.

[6] G. K. Savova, J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, K. C. Kipper-Schuler, and C. G. Chute, "Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications," *Journal of the American Medical Informatics Association*, vol. 17, no. 5, pp. 507–513, 2010.

[7] A. L. Beam, B. Kompa, A. Schmaltz, I. Fried, G. Weber, N. Palmer, X. Shi, T. Cai, and I. S. Kohane, "Clinical concept embeddings learned from massive sources of multimodal medical data," in *Pacific Symposium on Biocomputing 2020*. World Scientific, 2019, pp. 295–306.

[8] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in *Proceedings of the NAACL-HLT 2016*, 2016, pp. 1480–1489.

[9] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PloS one*, vol. 10, no. 7, 2015.

[10] M. Ancona, E. Ceolini, C. Öztireli, and M. Gross, "Towards better understanding of gradient-based attribution methods for deep neural networks," in *6th International Conference on Learning Representations (ICLR)*, no. 1711.06104. Arxiv-Computer Science, 2018, pp. 1–16.