LEARNING-BASED RADIOMIC PREDICTION OF TYPE 2 DIABETES MELLITUS USING IMAGE-DERIVED PHENOTYPES

Anonymous authors
Paper under double-blind review

ABSTRACT

Early diagnosis of Type 2 Diabetes Mellitus (T2DM) is crucial to enable timely therapeutic interventions and lifestyle modifications. As medical imaging data become more widely available for many patient populations, we sought to investigate whether image-derived phenotypic data could be leveraged in tabular learning classifier models to predict T2DM incidence without the use of invasive blood lab measurements. We show that both neural network and decision tree models that use image-derived phenotypes can predict patient T2DM status with recall scores as high as 87.6%. We also propose the novel use of these same architectures as 'SynthA1c encoders' that are able to output interpretable values mimicking blood hemoglobin A1C empirical lab measurements. Finally, we demonstrate that T2DM risk prediction model sensitivity to small perturbations in input vector components can be used to predict performance on covariates sampled from previously unseen patient populations.

1 INTRODUCTION

Type 2 Diabetes Mellitus (T2DM) is a chronic metabolic disorder that affects over 30 million patients in the United States and over 450 million patients worldwide (Khan et al., 2020; Xu et al., 2018). Patients diagnosed with T2DM are at an increased risk of comorbidities that include cardiovascular disease, retinopathy, peripheral neuropathy, and other complications (Albarakat and Guzu, 2019). Fortunately, early diagnosis and lifestyle interventions can prevent the onset of T2DM and are therefore critical to reduce the risk of disease progression and improve patient outcomes.

However, delayed clinical diagnosis of T2DM is frequent due to a low rate of screening. Polubriaginof et al. (2018) and Kaul et al. (2022) found that up to a third of patients do not receive T2DM screening as recommended by current national guidelines, and Porter et al. (2022) estimated that it would take over 24 hours per day for primary care physicians to follow national screening recommendations for every adult visit. Low rates of screening can further exacerbate healthcare disparities with respect to socioeconomic status and population marginalization.

Given these obstacles, machine learning has recently emerged as a promising tool to predict patient risk of T2DM and other diseases (Farran et al., 2013). However, current learning-based methods are limited to assessing disease risk from feature vectors derived from clinical biomarker and physical examination data (Liu et al., 2022; Joshi and Dhakal, 2021; Deberneh and Kim, 2021; Wu et al., 2021; Kopitar et al., 2020; Tigga and Garg, 2020; Sasar et al., 2017). While such models showcase promising predicting performance, they often lack real-world clinical applicability. This is because the blood biomarkers that are used to make a formal diagnosis of T2DM are trivially acquired alongside many of the blood biomarkers used to predict T2DM status.

Meanwhile, the usage of radiologic imaging in clinical medicine continues to increase every year (Dowhanik et al., 2022; Hong et al., 2020). For example, over 70 million computed tomography (CT) scans are performed annually (Smith-Bindman et al., 2019) and consequently offer a wealth of radiomic information that can potentially be used to estimate patient risk of T2DM as an incidental finding during an outpatient imaging appointment.

In this study, we sought to investigate whether radiomic metrics derived from clinically-acquired CT scans could be used to predict patient T2DM status. Our contributions are as follows:

1. To our knowledge, we are among the first to show that image-derived phenotypic data derived from clinical CT images can be synthesized with physical examination data to predict patient T2DM risk with accuracy comparable to existing methods.
2. To improve clinical interpretability of model outputs, we propose SynthA1c, a synthetic estimate of patient blood hemoglobin A1C (HbA1c) and therefore patient T2DM status.
3. We offer a novel model smoothness metric that we use to predict T2DM risk stratification performance on previously unseen, out-of-domain patient datasets.

The remaining sections are organized as follows: In Section 2, we introduce relevant technical and clinical background as well as related work in T2DM risk prediction. Section 3 details the dataset and data preparation procedure used in our experiments, and Section 4 outlines our experimental methods for model training and evaluation. We then offer our main results in Section 5, followed by a discussion on relevance and future work in Section 6.

2 Background and Related Work

2.1 Clinical Risk Factors for T2DM

One of the most important clinical laboratory values in the diagnosis of T2DM is a patient's hemoglobin A1C (HbA1c) level, which is a proxy for the level of blood sugar averaged over the past 3 months and is measured through a patient blood sample. Formally, a patient is considered *prediabetic* if their HbA1c is between 5.7% and 6.5% and *diabetic* if their HbA1c is greater than 6.5%. An HbA1c measurement under 5.7% is considered nondiabetic.

Fletcher et al. (2002) detail the clinical risk factors for T2DM, which include factors such as age, blood pressure, obesity, and self-reported race and gender. Such **clinically derived phenotypes** (CDPs) are features that can easily be obtained from patients or health records in most outpatient settings. Of note, a typical metric used for obesity quantification is the *body mass index* (BMI):

$$\text{Body Mass Index (BMI)} = \frac{\text{Weight [kg]}}{(\text{Height [m]})^2} \tag{1}$$

Other risk factors include elevated serum lipid and triglyceride levels, serum glucose levels, and markers of systemic inflammation. However, each of these measurements require a patient blood draw, during which an HbA1c test can readily be simultaneously conducted using the same blood sample. Therefore, while blood biomarkers may contribute to high T2DM predictive power reported in previous work (Liu et al., 2022; Deberneh and Kim, 2021; Wu et al., 2021; Kopitar et al., 2020), we argue that such models offer little utility in real-world clinical practice.

Other risk factors for T2DM do not require clinical laboratory tests, and may include diet, levels of physical activity, family history of metabolic disorders, and patient history of pregnancy-associated gestational diabetes (Fletcher et al., 2002). However, obtaining an accurate picture of these predictive features requires an in-depth and time-intensive patient interview, which may not be feasible in settings where an incidental calculation for T2DM risk estimation would be performed, such as during a radiology imaging visit. An ideal predictive model would seamlessly integrate into existing clinical workflows and require little additional information from the patient or their provider.

Finally, additional T2DM risk factors include central adiposity and nonalcoholic fatty liver disease (NAFLD), which is a condition that describes the buildup of excess fat in the liver. However, such features have been difficult to accurately quantify in outpatient settings, but can be easily estimated from CT scans in clinical practice. For example, liver fat buildup can be estimated by calculating the *spleen-hepatic attenuation difference* (SHAD) from an abdominal CT scan

$$\begin{aligned}&\text{Spleen-Hepatic Attenuation Difference (SHAD)}\\&\qquad= (\text{Spleen CT Attenuation [HU]}) - (\text{Liver CT Attenuation [HU]})\end{aligned} \tag{2}$$

Here, the liver (hepatic) CT attenuation and spleen CT attenuation are referred to as **image-derived phenotypes** (IDPs) estimated from segmentation analysis of patient CT scans and other imaging modalities.[1] Other IDPs, such as those derived from subcutaneous fat (SubQ Fat, the fat located just underneath the skin) and visceral fat (Visc Fat, the fat located between abdominal organs), can also be used to quantify central adiposity. Given the consistent increase in diagnostic imaging in recent years reported by Dowhanik et al. (2022), a T2DM risk prediction model could report estimated T2DM risk during an imaging visit from real-time analysis of CT scans and patient information. To our knowledge, such a model that incorporates IDP metrics has not been previously explored.

2.2 Related Work in Classification Methods

With regards to supervised learning methods, Uddin et al. (2019) found that support vector machines (SVMs), naïve Bayes classifiers, and random forest classifiers are explored most commonly in recent work on single disease prediction from patient phenotypic data, a type of tabular data. Decision tree-based models achieve promising prediction accuracy across explored tasks, and Chen and Guestrin (2016) demonstrated that incorporating scalable gradient boosting with forest classifiers, termed gradient-boosted decision trees (**GBDTs**), could yield state-of-the-art classification performance.

The consistent performance of GBDTs and derivative decision tree classifiers appears robust to the recent success of deep neural networks (DNNs). One of the most common DNN baselines is the fully-connected neural network (**FCNN**) (Zhang et al., 2017). More recently, Popov et al. (2019) introduced neural oblivious decision ensembles (**NODEs**), a class of DNNs that achieved classification performance on par with (and sometimes better than) decision tree models. The Feature Tokenizer + Transformer (**FT-Transformer**) proposed by Gorishniy et al. (2021) effectively adopts transformer architectures to tabular data, but still showed that there was no universally superior classifier when compared to decision trees and other algorithms. Arik and Pfister (2019) and Song et al. (2019) introduced attention-based tabular learning methods **TabNet** and **AutoInt**, respectively.

Clinically, a number of non-learning-based risk models based on patient questionnaires are used by physicians in outpatient settings, and include the FINDRISC score (Bernabe-Ortiz et al., 2018), ARIC score (Raynor et al., 2012), and QDScore (Collins and Altman, 2011) among others (Martinez-Millana et al., 2019; Bang et al., 2009). T2DM risk prediction has recently been explored in machine learning competitions and an increasing number of academic works. Liu et al. (2022); Wu et al. (2021); Joshi and Dhakal (2021); Deberneh and Kim (2021); Tigga and Garg (2020); and Kopitar et al. (2020) explored T2DM risk prediction in different patient populations, but limited their analysis to SVMs, decision trees, and baseline linear models and logistic regression. Furthermore, these models worked with feature vectors derived from patient physical exams and clinical lab values with no IDPs. To our knowledge, an extensive comparison between DNNs and GBDTs remains to be seen in T2DM risk classification tasks.

3 Dataset and Data Preparation

The data used for our study were made available by the Anonymized Institution BioBank (AIBB), an academic biobank established by the Anonymized Institution to advance the study of causes and treatments of a variety of diseases. The dataset consists of over 60K patients from the Anonymized Institution Health System mapped to over 10K ICD-9 codes. All patients provided informed consent to participate in the AIBB and to utilization of patient data, which was approved by the Institutional Review Board of the Anonymized Institution.

From the AIBB dataset, we obtained the following de-identified demographic and clinical variables: age, self-reported gender, self-reported race, height, weight, systolic and diastolic blood pressure, abdominal CT scans, and blood HbA1c measurements. Of note, the only clinical lab value extracted was HbA1c to be used as ground truth in model training and evaluation—no blood biomarkers were used as model inputs. Patients with any missing values were excluded from the dataset. We also restricted our analysis to only outpatient measurements.

Using the axial border classification network and visceral/subcutaneous fat segmentation network proposed and trained by MacLean et al. (2021), we estimated four IDPs from any given abdominal

[1] CT attenuation is the quantitative reduction in the intensity of X-rays as they travels through different parts of the body in CT imaging.

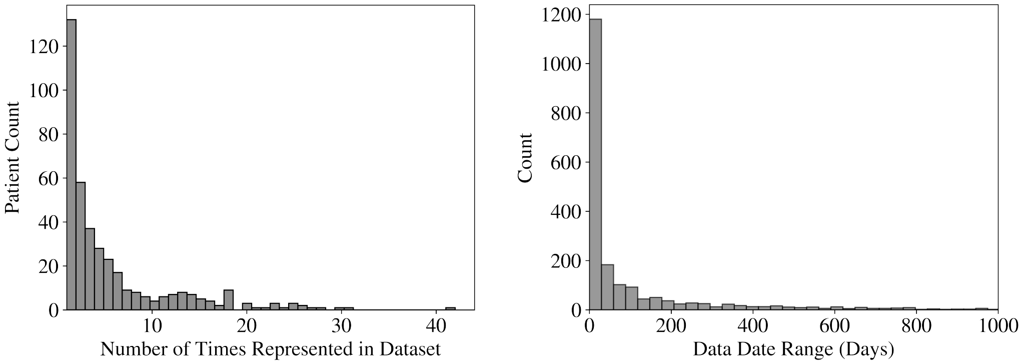

Figure 1: (Left) Distribution of the number of instances that a patient is represented in the dataset. (Right) Distribution of feature vector date ranges within our dataset.

CT study to be used as model inputs. These IDPs were (1) mean liver CT attenuation, (2) mean spleen CT attenuation, and estimated volume of (3) subcutaneous fat and (4) visceral fat.

The final pruned dataset consisted of 389 patients with a total of $N = 2077$ HbA1c measurements (1159 diabetic, 619 prediabetic, 299 nondiabetic).[2] 208 randomly selected samples were set aside as a holdout test set partition. Each HbA1c measurement was treated as a separate datapoint that could be used for model training or evaluation. To ensure that no patients were overrepresented in the dataset, we plotted a histogram of the number of times each patient was represented (Figure 1). The maximum frequency of 42 instances (1 occurence) only comprised about 2% of the overall dataset.

Aside from self-reported race and gender, which are assumed to be time-invariant features here, a patient i has a set $\{(x_{ij}, d_{ij})\}$ of any particular feature j within the dataset, where x_{ij} is a measurement of feature j for patient i and d_{ij} is the time at which x_{ij} was recorded. To construct a feature vector $\mathbf{x}$ associated with a particular HbA1c measurement y_i recorded on date $\tilde{d}_i$, we first define

$$(\hat{x}_{ij}, \hat{d}_{ij}) = \text{argmin}_{\{(x_{ij}, d_{ij})\}} \, ||d_{ij} - \tilde{d}_i||_1 \tag{3}$$

The feature vector $\mathbf{x}$ is formed by concatenating the selected measurements $\hat{x}_{ij}$ for all features j that minimize the time between the feature measurement and the HbA1c measurement y_i. We constructed feature vectors in this fashion for each of the $N = 2,077$ HbA1c measurements in our dataset. To ensure that the measurements in a given feature vector were made temporally close to one another, we plotted a histogram of daterange values, which we defined as the maximum length of time between any two features in $\mathbf{x}$ for each observation (Figure 1). The median daterange across our dataset was 18 days, and 87% of the feature vectors had a daterange no greater than 1 year.

4 Experimental Methods

Our features used as model inputs could be divided into two disjoint sets: clinically derived phenotypes (CDPs), which are derived from physical examination, and image-derived phenotypes (IDPs) that are estimated from abdominal CT scans herein. The specific CDPs and IDPs used depended on the model class—broadly, we explored two categories of models, which we refer to as either r-type and p-type (Table 1). r-type models were trained on '$r$aw' data types (i.e. height and weight, for instance), while p-type models were trained on '$p$rocessed' data types (i.e. BMI). Comparing the performance of r- and p- type models could help us better understand if using derivative processed metrics that are better clinically correlated with T2DM risk could yield better model performance.

In Section 5.1, we train fully supervised classifier models that learn to classify T2DM patient status from feature inputs. Ground truth was determined from HbA1c measurements based on the cutoff values presented in Section 2.1. Classifier models were trained using a binary cross entropy loss function and evaluated based on their recall, precision, specificity, and overall accuracy. To assess model performance for hyperparameter tuning, we maximized the accuracy score on our validation

[2] We stratify our dataset by self-reported gender and race in Table B1, and by age decade in Table B2.

partition consisting of 10% of the overall data. Final hyperparameter values are described in Table A1. The same set of hyperparameters were used to train an additional set of models to classify prediabetic status, where patients that are either prediabetic or diabetic according to their HbA1c lab value are grouped into a single category. This second task suffers from more class imbalance and therefore allows us to explore relative model performance on imbalanced datasets.

Table 1: Inputs for models trained on either raw data (r-Models) or a combination of raw and processed data (p-Models). Input data types are broken down into two categories: (1) Clinically Derived Phenotypes (CDPs) that can be assessed through a physical exam or patient interview, and (2) Image-Derived Phenotypes (IDPs) that are estimated from CT studies.

	Clinically Derived Phenotypes (CDPs)
r-Models	Race, Gender, Age, SBP, DBP, Height, Weight
p-Models	Race, Gender, Age, SBP, DBP, BMI
	Image-Derived Phenotypes (IDPs)
r-Models	SubQ Fat, Visc Fat, Liver CT Attenuation, Spleen CT Attenuation
p-Models	SubQ Fat, Visc Fat, Spleen-Hepatic Attenuation Difference (SHAD)

Separately in Section 5.2, we train fully supervised encoder models that learn to predict HbA1c values from feature inputs. These models were trained to minimize the L_2 loss relative to the patient's ground truth HbA1c lab measurement. The same loss function was also used for hyperparameter tuning through k=10-fold cross-validation from which the final values are reported in Table A3. All model training was performed on an Apple M1 MacBook Air using the CPU alone.

5 Results

5.1 Diabetes Mellitus Status Classification

Our results suggest that DNN and GBDT models perform comparably on both DM and DM/Pre-DM classification tasks (Table 2). While the NODE and FT-Transformer architectures achieved similar accuracy metrics, the FT-Transformer demonstrated the higher recall score, indicating a stronger ability to identify patients that may benefit from further workup for suspected diabetes (or pre-diabetes). Both the FT-Transformer and NODE models outperformed the AutoInt and baseline FCNN architectures with respect to overall accuracy. All DNN models outperformed the Zero-Rule and Weighted Random baselines, where the Zero-Rule classifier naïvely labels all patients as diabetic (or pre-diabetic), and the Weighted Random classifier randomly assigns patient diabetes status based on the proportion of positive samples in the training dataset. Furthermore, our learning-based models also outperformed the multi-rule classifier constructed from the prediabetes risk test from the American Diabetes Association (Bang et al., 2009).[3]

In comparing relative classification performance between GBDT and DNN models, it is difficult to identify a uniformly superior candidate. While the GBDT achieved the highest recall score for the DM/Pre-DM classification task, this performance came at the cost of lower precision and specificity metrics when compared to that of the NODE and FT-Transformer models.

5.2 SynthA1c Encoder: Predicting HbA1c Values

One of the primary limitations to widespread clinical adaptation of learning-based systems for healthcare is the inability for many models to intelligently work together with existing clinician workflows. Addressing this obstacle necessitates that model outputs be interpretable to be better understood by both healthcare providers and their patients. To this end, we sought to explore whether popular tabular learning architectures that performed well on T2DM classification tasks in Table 2 could simultaneously be used to encode a latent variable with values that closely approximate the clinical HbA1c lab measurement used to formally diagnose T2DM in clinical practice.

[3] Additional details on the Multi-Rule classifier are offered in Section A.1.

Table 2: Diabetes mellitus (DM) classification results using different classifier models that use both clinically derived phenotypes (CDPs) and image-derived phenotypes (IDPs) as input. Final hyperparameter values are included in Table A1. For comparison, our Multi-Rule classifier is a modified version of the official Prediabetes Risk Test provided by the American Diabetes Association and Centers for Disease Control and Prevention and predicts patient risk of diabetes through responses to a short clinical interview (Bang et al., 2009).

	DM Classification				DM + Pre-DM Classification			
Classifier	Recall	Precision	Specificity	Accuracy	Recall	Precision	Specificity	Accuracy
Zero-Rule	100	52.4	0.0	52.4	100	81.2	0.0	81.2
Weighted Random	44.4	75.5	41.0	43.8	87.2	81.4	12.8	73.4
Multi-Rule	67.0	54.9	39.4	53.8	92.9	83.5	20.5	79.3
r-FCNN	83.5	57.6	32.3	59.1	98.2	80.0	11.1	79.3
r-AutoInt	78.9	68.3	59.6	69.7	97.6	83.3	15.4	82.2
r-NODE	82.6	76.9	72.7	77.9	98.8	85.2	25.6	85.1
r-FT-Transformer	85.3	73.8	66.7	76.4	94.7	87.9	43.6	85.1
r-GBDT	87.2	76.6	70.7	79.3	95.3	88.0	43.6	85.6

We hypothesized that this synthetic HbA1c predictor, which we refer to as 'SynthA1c,' could be used to classify patient T2DM phenotype using traditional diabetic and pre-diabetic HbA1c cutoffs, providing output values that are better understood in clinical medicine. To our knowledge, we are the first to use covariates to predict HbA1c measurements as opposed to binary T2DM status.

Based on the results in Table 2, we focused our attention on two neural network models—NODE (Popov et al., 2019) and FT-Transformer (Gorishniy et al., 2021)—compared against GBDTs (Chen and Guestrin, 2016). Table 3 introduces the results of our encoder training using each of these three architectures. For each model, we calculated the root mean square error (RMSE) and Pearson correlation coefficient (PCC) between the predicted outputs and the ground truth. We then compared the predicted SynthA1c values with the traditional HbA1c threshold values for diabetes and pre-diabetes status classification to assess the utility of SynthA1c outputs in diagnosing T2DM.

Our results suggest that random forest encoders predicted SynthA1c values closest to ground truth HbA1c values according to both RMSE and PCC metrics. However, when SynthA1c values were then used for downstream T2DM status classification, we did not observe a uniformly superior model. The NODE and GBDT achieved similar recall scores while outperforming the FT-Transformer, but there were tradeoffs in model performance according to precision and specificity metrics. All models assessed performed better than the baseline ordinary least squares (OLS) encoder. Similar conclusions were made when using models that were trained on processed data types detailed in Table 1, such as BMI and spleen-hepatic attenuation difference. When feature vectors were expanded to the union of $r-$ and $p-$ model inputs (i.e. all of height, weight, and BMI were passed into a model, for instance), we observed no significant improvement in model performance.

Table 3: SynthA1c prediction results using different encoder models that use both clinically derived phenotypes (CDPs) and image-derived phenotypes (IDPs) as input. Final hyperparameter values are included in Table A3. r- (p-) prefixed models are fed raw (processed) inputs as outlined in Table 1.

			DM Classification			DM + Pre-DM Classification		
SynthA1c	RMSE (% A1C)	PCC	Recall	Precision	Specificity	Recall	Precision	Specificity
r-OLS	1.67	0.206	85.3	56.0	26.3	99.4	81.6	2.6
p-OLS	1.73	0.159	80.7	57.5	34.3	98.2	81.8	5.1
r-NODE	1.44	0.517	87.6	63.4	55.9	96.9	81.4	20.0
p-NODE	1.51	0.441	83.5	61.4	54.1	97.5	80.3	13.3
r-FT-Transformer	1.60	0.378	85.6	55.0	38.7	92.9	70.7	20.6
p-FT-Transformer	1.57	0.649	77.3	59.5	54.1	98.2	80.0	11.1
r-GBDT	1.36	0.567	87.2	66.4	51.5	96.4	82.3	10.3
p-GBDT	1.36	0.591	77.1	72.4	67.7	95.3	87.0	38.5

5.3 Characterizing Out-of-Domain Model Performance

Given the heterogeneity of patient populations, an important consideration in clinical applications of machine learning is the generalizability of T2DM classifiers to members of previously unseen patient groups. This issue can be addressed in part by utilizing representative datasets for model training and evaluation that more fully capture the diversity of phenotypic data encountered in the hospital. Indeed, unlike other T2DM datasets often cited in related work, such as the Pima Indians Diabetes Database (Smith et al., 1988) and a dataset limited to the Chinese elderly (Wu et al., 2021) that include data on only one particular gender and/or ethnic group, the AIBB-derived dataset used in our experiments includes a more diverse group of patients (Tables B1, B2). Nonetheless, our dataset is still affected by the geographic, environmental, and socioeconomic variables unique to the AIBB dataset patients that undoubtedly influence incidence of disease.

Prior work by Ng et al. (2022) and Jiang et al. (2021) have shown that model smoothness can be used to predict out-of-domain generalization and potential adversarial vulnerability of neural networks. However, these works largely limit their analysis to classifier networks. To evaluate SynthA1c encoder robustness, we wanted to develop an estimation of model manifold smoothness and define a corresponding smoothness metric $\mathbb{M}$ of our encoder models.

First, under the mild assumption that our SynthA1c encoder function $y : \mathbb{R}^{|\mathbf{x}|} \to \mathbb{R}$ is Lipschitz continuous, we can define a local manifold smoothness metric μ at $\mathbf{x} = \tilde{\mathbf{x}}$ given by

$$\begin{aligned} \mu(\tilde{\mathbf{x}}) &= \mathbb{E}_{\mathcal{N}(\tilde{\mathbf{x}})}\left[\frac{\sigma_y^{-1}||y(\mathbf{x}) - y(\tilde{\mathbf{x}})||_1}{||\delta\mathbf{x} \oslash \sigma_{\mathbf{x}}||_2}\right] \\ &= \left(\oint_{\mathcal{N}(\tilde{\mathbf{x}})\in\mathcal{D}} d\mathbf{x}\right)^{-1} \cdot \oint_{\mathcal{N}(\tilde{\mathbf{x}})\in\mathcal{D}} d\mathbf{x}\, \frac{\sigma_y^{-1}|y(\mathbf{x}) - y(\tilde{\mathbf{x}})|}{[(\delta\mathbf{x} \oslash \sigma_{\mathbf{x}})^T(\delta\mathbf{x} \oslash \sigma_{\mathbf{x}})]^{1/2}} \end{aligned} \tag{4}$$

where we have a feature vector $\tilde{\mathbf{x}}$ in domain $\mathcal{D}$ and a neighborhood $\mathcal{N}(\tilde{\mathbf{x}})$ around $\tilde{\mathbf{x}}$, and defining $\delta\mathbf{x} = \mathbf{x} - \tilde{\mathbf{x}}$, $\oslash$ as the Hadamard division operator, and $\sigma_{\mathbf{x}}$ as a vector of the estimated standard deviations of each feature over $\mathcal{D}$.[4] The exact expectation value over a given neighborhood $\mathcal{N}(\tilde{\mathbf{x}})$ is computationally intractable, but we can approximate it instead using Monte Carlo integration through an empirical sampling of $Q \gg 1$ random feature points $\mathbf{x}_k$ from $\mathcal{N}(\tilde{\mathbf{x}})$:

$$\mu(\tilde{\mathbf{x}}) \approx \frac{1}{Q}\sum_{k=1}^{Q} \frac{\sigma_y^{-1}|y(\mathbf{x}_k) - y(\tilde{\mathbf{x}})|}{[(\delta\mathbf{x}_k \oslash \sigma_{\mathbf{x}})^T(\delta\mathbf{x}_k \oslash \sigma_{\mathbf{x}})]^{1/2}} \tag{5}$$

Using this expression for μ, we can define a metric $\mathbb{M}$ for the global encoder manifold smoothness over a domain $\mathcal{D}$. Here, we propose $\mathbb{M}$ as the expectation value of $\mu(\mathbf{x})$ over $\mathcal{D}$, which can again be similarly approximated by an empirical sampling of N feature vectors $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\} \in \mathcal{D}$.

We hypothesized that the global smoothness metric $\mathbb{M}$ inversely correlates with model performance on an out-of-domain dataset. To evaluate this experimentally, we assessed model performance on two previously unseen T2DM datasets: (1) the Iraqi Medical City Hospital dataset (Rashid, 2020); and (2) the AIBB inpatient dataset. The Iraqi dataset contains data on 1,000 patients with age, gender, BMI, and HbA1c measurements. Because of the limited feature set, we trained additional SynthA1c encoders on the AIBB outpatient dataset using only these features (referred to as p'-type models). The AIBB inpatient dataset consists of 2,066 inpatient measurements of the same datatypes as the outpatient dataset described in Section 3. Generally speaking, inpatients are required to stay overnight in the hospital and can therefore be more unwell.

Based on the results on the Iraqi Medical Center Dataset in Table 4, we observed that as the global smoothness metric $\mathbb{M}$ decreases across the three evaluated models, corresponding to a more globally smooth model, the RMSE in SynthA1c prediction decreases and the PCC increases, which corresponds to better predictive performance on the out-of-domain dataset. This supports our initial hypothesis that smoother models may generalize better to unseen datasets. We also noted larger

[4] Broadly, we can think of $\mu(\tilde{\mathbf{x}})$ as the ratio of deviations in the normalized model output to the Euclidean norm of deviations in the normalized covariates over a neighborhod $\mathcal{N}(\tilde{\mathbf{x}})$ around $\mathbf{x} = \tilde{\mathbf{x}}$. Note that because the volume integral over $\mathcal{N}(\tilde{\mathbf{x}})$ in Equation 4 is cancelled out in the Monte Carlo approximation in Equation 5, μ cannot be compared across feature sets with different cardinalities.

Table 4: Diabetes mellitus (DM) encoder model sensitivity and out-of-domain generalization results. Global smoothness metric values $\mathbb{M}$ were evaluated on the AIBB outpatient dataset and are all multiplied by 10^2 here for legibility. Model inference was evaluated on the Iraqi Medical Center Dataset and the AIBB Inpatient dataset. Because the Iraqi dataset does not include blood pressure measurements or imaging data, we trained additional encoder models on CDPs only using a separate set of features p'—these p'-type, CDP-only models are trained on all p-type model inputs as in Table 1 except systolic and diastolic blood pressure. We used the same hyperparameters in training p-type models for p'-type models, as delineated in Table A3.

CDP Training Only (Race, Gender, Age, and BMI)						**CDP + IDP Training**			
		Iraqi Dataset		**AIBB Inpatient**				**AIBB Inpatient**	
SynthA1c Encoder	$\mathbb{M}$	RMSE (% A1C)	PCC	RMSE (% A1C)	PCC	SynthA1c Encoder	$\mathbb{M}$	RMSE (% A1C)	PCC
p'-NODE	1.43	3.62	0.154	1.76	0.512	r-NODE	28.3	1.23	0.795
p'-FTTransformer	1.07	3.04	0.246	1.90	0.331	r-FTTransformer	23.2	1.58	0.617
p'-GBDT	3.28	6.25	0.021	1.54	0.674	r-GBDT	37.3	1.12	0.823

RMSE error values using the Iraqi Medical Center Dataset when compared to the AIBB outpatient test dataset results in Table 3. Interestingly, we found that this relationship did not hold when considering the AIBB inpatient dataset; in fact, model predictive performance was *inversely* correlated with global smoothness. This could suggest that the AIBB inpatient and outpatient dataset distributions are more similar than initially predicted, and that inpatient T2DM disease phenotypes are not substantially distinct from outpatient disease within an otherwise identical patient population.

To further investigate this hypothesis, we computed the empirical Kullback-Leibler (KL) divergence between each of the test dataset distributions and the training dataset distribution with respect to the CDP features available in all our datasets: race, gender, age, BMI, and HbA1c. We assumed that our training outpatient dataset is drawn from an unknown distribution $\mathcal{Q}(\mathbf{x})$, and each of our test datasets—the AIBB outpatient, AIBB inpatient, and Iraqi Medical Center datasets—are separately drawn from unknown distributions $\mathcal{P}_{\text{Outpatient}}(\mathbf{x}), \mathcal{P}_{\text{Inpatient}}(\mathbf{x}), \mathcal{P}_{\text{Iraqi}}(\mathbf{x})$, respectively. Our results in Table 5 show that, as expected, both AIBB-derived datasets share a more similar distribution to the AIBB-derived outpatient training dataset when compared to the separate Iraqi dataset. This correlates with relative model performance in Tables 3 and 4. Interestingly, we also found that the KL divergence between the inpatient test set and training dataset was *lower* than that between the outpatient test set and training dataset. This agrees with the lower RMSE metric reported on the inpatient dataset in Table 4 when compared to that on the outpatient dataset in Table 3.

Table 5: Empirical KL Divergences comparing different test datasets with the training dataset $\mathcal{Q}$.

$D_{KL}(\mathcal{P}_{\text{Outpatient}}\|\|\mathcal{Q})$	$D_{KL}(\mathcal{P}_{\text{Inpatient}}\|\|\mathcal{Q})$	$D_{KL}(\mathcal{P}_{\text{Iraqi}}\|\|\mathcal{Q})$
1.84	0.227	31.2

Qualitatively, our results could be explained by a greater degree of homogeneity within the inpatient population that is better captured by the features observed by the encoder model during training. Additional discussion is offered in Section B.3. Further work is warranted to validate the proposed utility of the proposed metric $\mathbb{M}$ across various tasks within the broader machine learning community, which is outside the scope of our discussion herein.

5.4 Ablation Studies

Which input features are most important for classification performance? Until now, prior T2DM classifiers have used only blood lab measurements and physical examination data to predict T2DM status. In contrast, our classifiers presented herein are the first to incorporate IDPs as input model features. To better understand the benefit of using IDPs in conjunction with CDPs, we evaluated classifier performance on models trained using either only CDPs or only IDPs and compared them to corresponding models trained using both CDPs and IDPs as inputs.

Our results suggest that while classifier models trained only on CDPs generally outperform those trained only on IDPs, the best performance is achieved when combining CDPs and IDPs together

(Table 6). This further validates the clinical utility of incorporating IDPs into patient diagnosis and disease risk stratification first proposed by MacLean et al. (2021) and related work.

Table 6: Ablation study assessing model performance as a function of whether clinically derived phenotypes (CDPs) and/or image-derived phenotypes (IDPs) are passed in as inputs. Here, we assessed performance for the three subjectively best performing classifiers from Table 2: NODE, FT-Transformer, and GBDT. Final hyperparameter values are included in Table A4.

	DM Classification				DM + Pre-DM Classification			
r-**NODE**	Recall	Precision	Specificity	Accuracy	Recall	Precision	Specificity	Accuracy
CDPs Only	77.1	73.7	69.7	73.5	95.9	86.6	35.9	84.6
IDPs Only	73.4	76.9	75.8	74.5	90.5	90.5	59.0	84.6
CDPs + IDPs	82.6	76.9	72.7	77.9	98.8	85.2	25.6	85.1
r-**FT-Transformer**	Recall	Precision	Specificity	Accuracy	Recall	Precision	Specificity	Accuracy
CDPs Only	78.0	76.6	73.7	75.9	95.3	85.6	30.8	83.2
IDPs Only	71.6	60.5	48.5	60.6	92.9	85.8	33.3	81.7
CDPs + IDPs	85.3	73.8	66.7	76.4	94.7	87.9	43.6	85.1
r-**GBDT**	Recall	Precision	Specificity	Accuracy	Recall	Precision	Specificity	Accuracy
CDPs Only	80.7	68.6	59.6	70.7	98.8	84.8	23.1	84.6
IDPs Only	73.4	75.5	73.7	73.6	96.4	87.6	41.0	86.0
CDPs + IDPs	87.2	76.6	70.7	79.3	95.3	88.0	43.6	85.6

6 Discussion and Conclusion

Our work highlights the value of integrating IDPs extracted using AI-based radiomic analysis into both neural network and decision tree models for predicting T2DM patient phenotypes. We demonstrated that fully supervised models that utilize IDPs and CDPs together can accurately predict T2DM status as a potential incidental finding during radiologic imaging appointments, requiring minimal additional information from patients and healthcare providers. Simultaneously, we showed that popular tabular learning architectures could act as novel SynthA1c encoders to predict HbA1c measurements noninvasively and ultimately improve the clinical interpretability of model outputs. Finally, we demonstrate that model sensitivity to perturbations in input feature vectors may be correlated with prediction performance on previously unseen data sampled from out-of-domain patient populations, although additional validation studies on separate tasks are needed.

Future work remains to be done to improve the predictive power of learning-based patient T2DM status prediction. Firstly, the increasingly widespread availability of clinical imaging data enables potential work in tracking changes in patient-specific imaging studies and extracted IDPs over periods of time, which could be indicative of disease progression or remission. Such time series data could represent powerful additional features that better take into the account the longitudinal effects of therapeutic regimens, lifestyle modifications, and other factors. For example, prior work by Pimentel et al. (2018) showed that time series analysis using physical exam features alone could achieve comparable predictive power compared to corresponding models trained on static feature vectors consisting of both physical exam and more invasive clinical lab measurements.

Medical biobanks, such as the Anonymized Institution BioBank used to construct our dataset herein, also feature other information-rich data types in addition to imaging and clinical data. Modern health datasets now include genomic sequencing, wearable devices, and histological slides among many others that can enable powerful advancements in precision medicine in addition to disease and prognosis forecasting. As such data become increasingly available, multimodal ML methods that are able to effectively incorporate all these data types may improve upon current techniques.

Code and Data Availability

The code for this project is available at github.com/anonymized-user/RepositoryName and is licensed under the MIT License. The AIBB dataset is available to investigators upon reasonable request.

Ethics Statement

The results presented herein are on datasets that include real patient information and medical data from the Anonymized Institution Health System and from the Iraqi Medical City Hospital. All patients in the Anonymized Institution BioBank provided informed consent to the utilization of individual patient data, and corresponding research was approved by the Institutional Review Board of the Anonymized Institution. The Iraqi Medical City Hospital dataset was collected from the Iraqi society, and its curation is described by Rashid (2020). It is made publicly available to all researchers by the dataset authors at DOI 10.17632/wj9rwkp9c2.1.

As introduced in Section 5.3, we emphasize the importance of training predictive models on representative patient datasets, and the results herein should not be immediately taken to generalize perfectly to other real-world datasets in high-stakes medical applications. Furthermore, risk stratification models, such as the ones reported herein, should be used in conjunction with advice from medical experts and should not be interpreted without consulting a healthcare provider. Additional work, such as in conformal prediction (Stutz et al., 2021; Bastani et al., 2022) and prediction sets (Park et al., 2022), should also be considered to better quantify uncertainty in model predictions for disease diagnosis applications. We hope that our work will have a positive impact in diabetes screening and both patient and population health.

Reproducibility Statement

Section 3 includes a detailed discussion of our dataset preparation, and Section 4 explicitly delineates dataset partitions, model inputs, and experimental training procedures. Hyperparameters used during model training are reported in Section A. We also make our Python implementation for all experiments reported herein available at `github.com/anonymized-user/RepositoryName`.

References

Moien Abdul Basith Khan, Muhammad Jawad Hashim, Jeffrey Kwan King, Romona Devi Govender, Halla Mustafa, and Juma Al Kaabi. Epidemiology of Type 2 Diabetes - global burden of disease and forecasted trends. *J Epidemiol Glob Health*, 10:107–111, 2020. doi: 10.2991/jegh.k.191028.001.

Guifeng Xu, Buyun Liu, Yangbo Sun, Yang Du, Linda G Snetselaar, Frank B Hu, and Wei Bao. Prevalence of diagnosed type 1 and type 2 diabetes among US adults in 2016 and 2017: population based study. *BMJ*, 362, 2018. doi: 10.1136/bmj.k1497.

Mohammed Albarakat and Ali Guzu. Prevalence of type 2 diabetes and their complications among home health care patients at Al-Kharj military industries corporation hospital. *J Family Med Prim Care*, 8:3303–3312, 2019. doi: 10.4103/jfmpc.jfmpc_634_19.

Fernanda C G Polubriaginof, Ning Shang, George Hripcsak, Nicholas P Tatonetti, and David K Vawdrey. Low screening rates for diabetes mellitus among family members of affected relatives. *AMIA Annu Symp Proc*, pages 1471–1477, 2018.

Padma Kaul, Luan Manh Chu, Doughlas C Dover, Roseanne O Yeung, Dean T Eurich, and Sonia Butalia. Disparities in adherence to diabetes screening guidelines among males and females in a universal care setting: A population-based study of 1,380,697 adults. *Lancet Regional Health*, 2022. doi: 10.1016/j.lana.2022.100320.

Justin Porter, Cynthia Boyd, M Reza Skandari, and Neda Laiteerapong. Revisiting the time needed to provide adult primary care. *J Gen Intern Med*, 2022. doi: 10.1007/s11606-022-07707-x.

Bassam Farran, Arshad Mohamed Channanath, Kazem Behbehani, and Thangavel A Thanaraj. Predictive models to assess risk of type 2 diabetes, hypertension and comorbidity: machine-learning algorithms and validation using national health data from kuwait—a cohort study. *BMJ Open*, 3 (5), 2013. doi: 10.1136/bmjopen-2012-002457.

Qing Liu, Miao Zhang, Yifang He, Lei Zhang, Jingui Zhou, Yaqiong Yan, and Yan Guo. Predicting the risk of incident type 2 diabetes mellitus in chinese elderly using machine learning techniques. *J Pers Med*, 12:905, 2022. doi: 10.3390/jpm12060905.

Ram D Joshi and Chandra K Dhakal. Predicting type 2 diabetes using logistic regression and machine learning approaches. *Int J Environ Res Public Health*, 18:7346, 2021. doi: 10.3390/ijerph18147346.

Henock M Deberneh and Intaek Kim. Prediction of type 2 diabetes based on machine learning algorithm. *Int J Environ Res Public Health*, 18:3317, 2021. doi: 10.3390/ijerph18063317.

Yang Wu, Haofei Hu, Jinlin Cai, Runtian Chen, Xin Zuo, Heng Cheng, and Dewen Yan. Machine learning for predicting the 3-year risk of incident diabetes in Chinese adults. *Frontiers in Public Health*, 9, 2021. doi: 10.3389/fpubh.2021.626331.

Leon Kopitar, Primov Kocbek, Leona Cilar, Aziz Sheikh, and Gregor Stiglic. Early detection of type 2 diabetes mellitus using machine learning-based prediction models. *Nat Sci Rep*, 2020. doi: 10.1038/s41598-020-68771-z.

Neha Prerna Tigga and Shruti Garg. Prediction of type 2 diabetes using machine learning classification methods. *Proceia Computer Science*, 167:706–716, 2020. doi: 10.1016/j.procs.2020.03.336.

Ali Sasar, Osman Ozkaraca, Musa Peker, and Gurbuz Akcay. Estimation of HbA1c value using artificial neural networks. *Glob J Comp Sci*, 7:1–7, 2017. doi: 10.18844/gjcs.v7i1.2691.

Sebastian P D Dowhanik, Nicola Schieda, Michael N. Patlas, Fateme Salehi, and Christian B van der Pol. Doing more with less: CT and MRI utilization in Canada 2003–2019. *Canadian Association of Radiologists Journal*, 73(3):592–594, 2022. doi: 10.1177/08465371211052012.

Arthur S Hong, David Levin, Laurence Parker, Vijay M Rao, Dennis Ross-Degnan, and J Frank Wharam. Trends in diagnostic imaging utilization among Medicare and commercially insured adults from 2003 through 2016. *Radiology*, 294(2):342–350, 2020. doi: 10.1148/radiol.2019191116.

Rebecca Smith-Bindman, Marilyn L. Kwan, Emily C. Marlow, Mary Kay Theis, Wesley Bolch, Stephanie Y. Cheng, Erin J. A. Bowles, James R. Duncan, Robert T. Greenlee, Lawrence H. Kushi, Jason D. Pole, Alanna K. Rahm, Natasha K. Stout, Sheila Weinmann, and Diana L. Miglioretti. Trends in use of medical imaging in US health care systems and in Ontario, Canada, 2000-2016. *JAMA*, 322(9):843–856, 2019. doi: 10.1001/jama.2019.11456.

Barbara Fletcher, Meg Gulanick, and Cindy Lamendola. Risk factors for Type 2 Diabetes Mellitus. *J Card Nurs*, 16:17–23, 2002.

Shahadat Uddin, Arid Khan, Md Ekramul Hossain, and Mohammad Ali Moni. Comparing different supervised machine learning algorithms for disease prediction. *BMC Med Inform Decis Mak*, 19 (281), 2019. doi: 10.1186/s12911-019-1004-8.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proc ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining*, page 785–794, 2016. doi: 10.1145/2939672.2939785.

Yuchen Zhang, Jason Lee, Martin Wainwright, and Michael I Jordan. On the learnability of fully-connected neural networks. In *Proc Int Conf AI and Statistics*, volume 54, pages 83–91, 2017.

Sergei Popov, Stanislav Morozov, and Artem Babenko. Neural oblivious decision ensembles for deep learning on tabular data, 2019.

Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Revisiting deep learning models for tabular data, 2021.

Sercan O Arik and Tomas Pfister. TabNet: Attentive interpretable tabular learning, 2019.

Weiping Song, Chence Shi, Zhiping Xiao, Zhijian Duan, Yewen Xu, Ming Zhang, and Jian Tang. AutoInt: Automatic feature interaction learning via self-attentive neural networks. In *Proc ACM Int Conf on Information and Knowledge Management*, 2019. doi: 10.1145/3357384.3357925.

Antonio Bernabe-Ortiz, Pablo Perel, Juan Jaime Miranda, and Liam Smeeth. Diagnostic accuracy of the Finnish diabetes risk score (FINDRISC) for undiagnosed T2DM in Peruvian population. *Prim Care Diabetes*, 12, 2018. doi: 10.1016/j.pcd.2018.07.015.

L. A. Raynor, James S. Pankow, Bruce B. Duncan, Maria I. Schmidt, Ron C. Hoogeveen, Mark A. Pereira, J. Hunter Young, and Christie M. Ballantyne. Novel risk factors and the prediction of type 2 diabetes in the atherosclerosis risk in communities (ARIC) study. *Diabetes Care*, 36(1): 70–76, 2012. doi: 10.2337/dc12-0609.

G S Collins and Douglas G Altman. External validation of QDSCORE for predicting the 10-year risk of developing type 2 diabetes. *Diabet Med*, 28:599–607, 2011. doi: 10.1111/j.1464-5491.2011.03237.x.

Antonia Martinez-Millana, María Argente-Pla, Bernardo Valdivieso Martinez, Vicente Traver Salcedo, and Juan Francisco Merino-Torres. Driving type 2 diabetes risk scores into clinical practice: Performance analysis in hospital settings. *J Clin Med*, 8:107, 2019. doi: 10.3390/jcm8010107.

Heejung Bang, Alison M Edwards, Andrew S Bomback, Christie M Ballantyne, David Brillon, Mark A Callahan, Steven M Teutsch, Alvin I Mushlin, and Lisa M Kern. Development and validation of a patient self-assessment score for diabetes risk. *Ann Intern Med*, 151:775–83, 2009. doi: 10.7326/0003-4819-151-11-200912010-00005.

Matthew T MacLean, Qasim Jehangir, Marijana Vujkovic, Yi-An Ko, Harold Litt, Arijitt Borthakur, Hersh Sagreiya, Mark Rosen, David A Mankoff, Mitchell D Schnall, Haochang Shou, Julio Chirinos, Scott M Damrauer, Drew A Torigian, Rotonya Carr, Daniel J Rader, and Walter R Witschey. Quantification of abdominal fat from computed tomography using deep learning and its association with electronic health records in an academic biobank. *J Am Med Inform Assoc*, 28: 1178–1187, 2021. doi: 10.1093/jamia/ocaa342.

Jack W Smith, J E Everhart, W C Dickson, W C Knowler, and R S Johannes. Using the ADAP learning algorithm to forecast the onset of Diabetes Mellitus. *Proc Sym Comp App and Med Care*, pages 261–265, 1988.

Nathan Ng, Neha Hulkund, Kyunghyun Cho, and Marzyeh Ghassemi. Predicting out-of-domain generalization with local manifold smoothness, 2022.

Zijian Jiang, Jianwen Zhou, and Haiping Huang. Relationship between manifold smoothness and adversarial vulnerability in deep learning with local errors. *Chinese Physics B*, 30(4), 2021. doi: 10.1088/1674-1056/abd68e.

Ahlam Rashid. Iraqi diabetes dataset, 2020. URL `data.mendeley.com/datasets/wj9rwkp9c2/1`.

Angela Pimentel, André V Carreiro, Rogério T Ribeiro, and Hugo Gamboa. Screening diabetes mellitus 2 based on electronic health records using temporal features. *Health Informatics J*, 24: 194–205, 2018. doi: 10.1177/1460458216663023.

David Stutz, Dvijotham Krishnamurthy, Ali Taylan Cemgil, and Arnaud Doucet. Learning optimal conformal classifiers, 2021.

Osbert Bastani, Varun Gupta, Christopher Jung, Georgy Noarov, Ramya Ramalingam, and Aaron Roth. Practical adversarial multivalid conformal prediction, 2022.

Sangdon Park, Edgar Dobriban, Insup Lee, and Osbert Bastani. PAC prediction sets for meta-learning, 2022.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014.

Patrick S Kamath, Russell H Wiesner, Michael Malinchoc, Walter Kremers, Terry M Therneau, Catherine . Kosberg, Gennaro D'Amico, E Rolland Dickson, and W Ray Kim. A model to predict survival in patients with end-stage liver disease. *Hepatology*, 33(2):464–470, 2001. doi: 10.1053/jhep.2001.22172.

Stuart J Pocock, Cono A Ariti, John J V McMurray, Aldo Maggioni, Lars Kober, Iain B Squire, Karl Swedberg, Joanna Dobson, Katrina K Poppe, Gillian A Whalley, and Rob N Doughty. Predicting survival in heart failure: a risk score based on 39,372 patients from 30 studies. *European Heart Journal*, 34(19):1404–1413, 2012. doi: 10.1093/eurheartj/ehs337. URL `https://doi.org/10.1093/eurheartj/ehs337`.

Graham Teasdale and Bryan Jennett. Assessment of coma and impaired consciousness. A practical scale. 304:81–84, 1974. doi: 10.1016/s0140-6736(74)91639-0.

A IMPLEMENTATION DETAILS

A.1 T2DM CLASSIFIER EXPERIMENTS

Table A1 summarizes the final tunable hyperparameter values used for our T2DM classifier experiments introduced in Section 5.1.

Table A1: Hyperparameters for the T2DM status classification experiment. All non-GBDT models were trained using an Adam optimizer with default parameters $\beta_1 = 0.9, \beta_2 = 0.999$ (Kingma and Ba, 2014).

Hyperparameter	r-**FCNN**	r-**AutoInt**	r-**NODE**	r-**FT-Transformer**	r-**GBDT**
Number of Epochs/Boosted Trees	150	150	150	150	32
η Learning Rate (LR)	0.01	0.001	0.03	0.0001	0.1
LR Step Size (Epochs)	25	50	100	100	—
γ LR Decay Rate	0.5	0.1	0.5	0.5	—
Batch Size	128	128	64	64	—
Dropout	0.0	0.01	0.0	0.01	—
Activation	ReLU	ReLU	ReLU	ReLU	—
Maximum Tree Depth	—	—	—	—	16
α L_1 Regularization	—	—	—	—	1
λ L_2 Regularization	—	—	—	—	2
Early Stopping	No	No	No	No	No

In our results in Table 2, we compare our learning-based models against a multi-rule classifier inspired by a similar classifier provided through a collaboration by the American Diabetes Association (ADA) and Centers for Disease Control and Prevention (CDC). The official ADA/CDC classifier from Bang et al. (2009) predicts patient risk of developing prediabetes through assigning points to specific responses to a short clinical interview. A score greater than a pre-defined threshold corresponds to increased risk for prediabetes and ultimately T2DM. Similar score-based metrics are used clinically for a wide variety of different diseases (Kamath et al., 2001; Pocock et al., 2012; Teasdale and Jennett, 1974; Bernabe-Ortiz et al., 2018). However, because some of the Bang et al. (2009) classifier questions ask for patient information that we do not consider here, such as family history and past pregnancy information, we only selected a subset of questions from the ADA/CDC classifier to use in our baseline multi-rule classifier for Table 2, which we detail in Table A2.

Because our multi-rule classifier does not use the same exact set of questions as the ADA/CDC classifier, the ADA/CDC threshold scores to classify patients as having increased risk for T2DM is not valid for us to use here. Therefore, based on analyzing the distribution of scores for diabetic, prediabetic, and non-diabetic patients, we chose to empirically set the threshold cutoff for prediabetes as having a score of 3 or higher from Table A2, and the cutoff for diabetes as a score of 5 or higher. For comparison, the cutoff for the classifier proposed by Bang et al. (2009) was a score of 5 or higher for prediabetic status.

A.2 SYNTHA1C ENCODER EXPERIMENTS

Table A3 summarizes the final tunable hyperparameter values used for our SynthA1c encoder experiments introduced in Section 5.2.

Table A2: **Modified ADA/CDC Multi-Rule Classifier**: Questionnaire used for baseline classification in Table 2. All questions and point assignments are the same as those used in the official ADA/CDC diabetes risk test (Bang et al., 2009). Due to lack of data, we excluded questions from the official test that asked about (1) gestational diabetes diagnosis, (2) family history of diabetes, and (3) exercise activity level.

Question	**Points**
1. How old are you?	
$<$40 years old	0
40-49 years old	1
50-59 years old	2
$>$60 years old	3
2. Do you identify as a man or a woman?	
Woman	0
Man	1
3. Do you have a systolic blood pressure $>$130 mmHg and/or a diastolic blood pressure $>$80 mmHg?	
No	0
Yes	1
4. What is your weight category? Weight categories are defined originally by Bang et al. (2009).	
Not overweight or obese	0
Overweight	1
Obese	2
Extremely obese	3
Sum all points to calculate final score.	

Table A3: Hyperparameters for the SynthA1c experiment. All non-GBDT models were trained using an Adam optimizer with default parameters $\beta_1 = 0.9, \beta_2 = 0.999$ (Kingma and Ba, 2014).

Hyperparameter	r-**NODE**	p-**NODE**	r-**FT-Transformer**	p-**FT-Transformer**	r-**GBDT**	p-**GBDT**
Number of Epochs/Boosted Trees	100	100	100	100	32	32
η Learning Rate (LR)	0.01	0.03	0.00001	0.001	0.25	0.1
LR Step Size (Epochs)	40	40	—	50	—	—
γ LR Decay Rate	0.5	0.5	—	0.5	—	—
Batch Size	16	16	4	128	—	—
Dropout	0.0	0.0	0.02	0.01	—	—
Activation	ReLU	ReLU	ReLU	ReLU	—	—
Maximum Tree Depth	—	—	—	—	6	8
α L_1 Regularization	—	—	—	—	0	1
λ L_2 Regularization	—	—	—	—	1	4
Early Stopping	No	No	No	No	No	No

A.3 Ablation Study Experiments

Table A4 summarizes the final tunable hyperparameter values used for our classifier ablation experiments discussed in Section 5.4.

B Additional Results

A breakdown by demographic features of our AIBB outpatient training dataset used for model training is reported in Tables B1 and B2.

B.1 Diabetes Mellitus Status Classification Experiments

Receiver operating characteristic (ROC) curves for the non-baseline classifiers delineated in Table 2 are shown in Figure B1.

Table A4: Hyperparameters for comparing models trained on only CDPs or only IDPs with those trained using a combination of the two. The final hyperparameter values for models trained using both CDPs and IDPs are included in Table A1. All non-GBDT models were trained using an Adam optimizer with default parameters $\beta_1 = 0.9, \beta_2 = 0.999$ (Kingma and Ba, 2014).

	CDPs Only			**IDPs Only**		
Hyperparameter	r-**NODE**	r-**FT-Transformer**	r-**GBDT**	r-**NODE**	r-**FT-Transformer**	r-**GBDT**
Number of Epochs/Boosted Trees	150	150	32	150	150	32
η Learning Rate (LR)	0.03	0.001	0.1	0.03	0.001	0.3
LR Step Size (Epochs)	100	50	—	100	100	—
γ LR Decay Rate	0.5	0.5	—	0.5	0.5	—
Batch Size	16	128	—	64	32	—
Dropout	0.0	0.01	—	0.0	0.01	—
Activation	ReLU	ReLU	—	ReLU	ReLU	—
Maximum Tree Depth	—	—	16	—	—	16
α L_1 Regularization	—	—	2	—	—	2
λ L_2 Regularization	—	—	4	—	—	4
Early Stopping	No	No	No	No	No	No

Table B1: Dataset stratified by self-reported gender and self-reported race.

	White	Hispanic	Black	Asian	Pacific Islander	Native American	Other/Unknown	Total
Male	401	26	427	8	6	0	12	880
Female	319	14	821	28	0	5	10	1197
Total	720	40	1248	36	6	5	22	**2077**

B.2 SynthA1c Encoder Experiments

To further interrogate our SynthA1c encoders and investigate the possibility of biased performance, we examined whether model performance varied as a function of demographic features such as self-reported gender and ethnicity (Figures B2, B3). Focusing our attention on our GBDT encoders, the best performing architecture based on test RMSE and PCC metrics, we observed no qualitative difference in model performance when results were stratified by either self-reported gender or race.

B.3 Out-of-Domain Model Performance Experiments

Figure B4 showcases the pairwise relationships between BMI, age, and Hemoglobin A1c measurements across three different datasets in order to better understand the domain characterization results from Tables 4 and 5.

Table B2: Dataset stratified by age.

Age Decade	20-29	30-39	40-49	50-59	60-69	70-79	80-89	Total
Count	31	89	362	593	680	299	23	**2077**

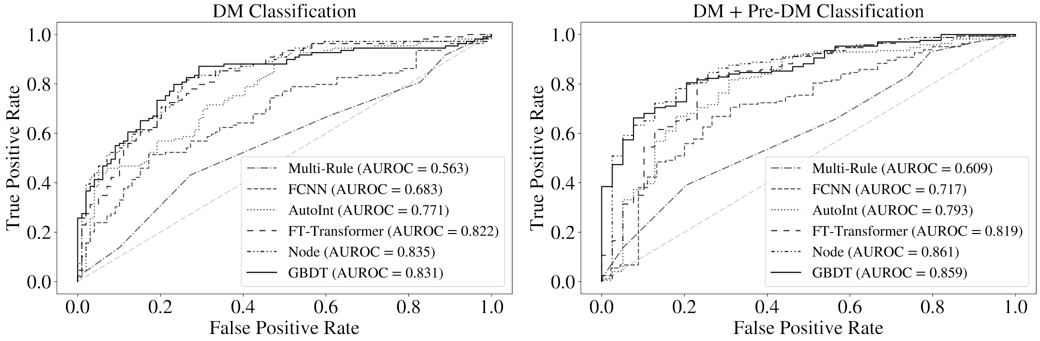

Figure B1: Receiver operating characteristic (ROC) curves for diabetic classification (left) and diabetic/pre-diabetic classification (right) tasks. Area under the ROC (AUROC) values are included for each classifier in the legend.

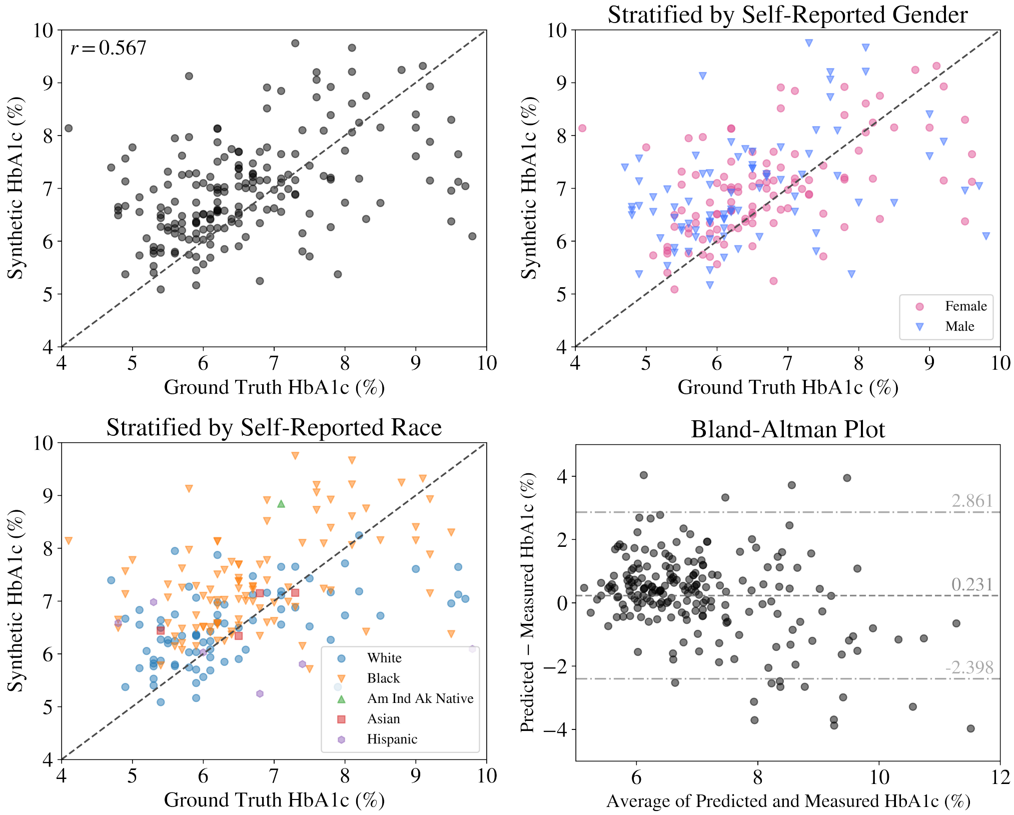

Figure B2: Quantitative analysis of the r-GBDT SynthA1c encoder described in Table 3. (**Top left**) Scatter plot of SynthA1c model outputs vs ground truth HbA1c lab measurements for $N = 208$ test datapoints. (**Top right**) Same plot stratified by self-reported gender. (**Bottom left**) Same plot stratified by self-reported race. (**Bottom right**) Bland-Altman plot comparing SynthA1c and HbA1c values. Figure B3 includes a similar panel for the p-GBDT SynthA1c encoder.

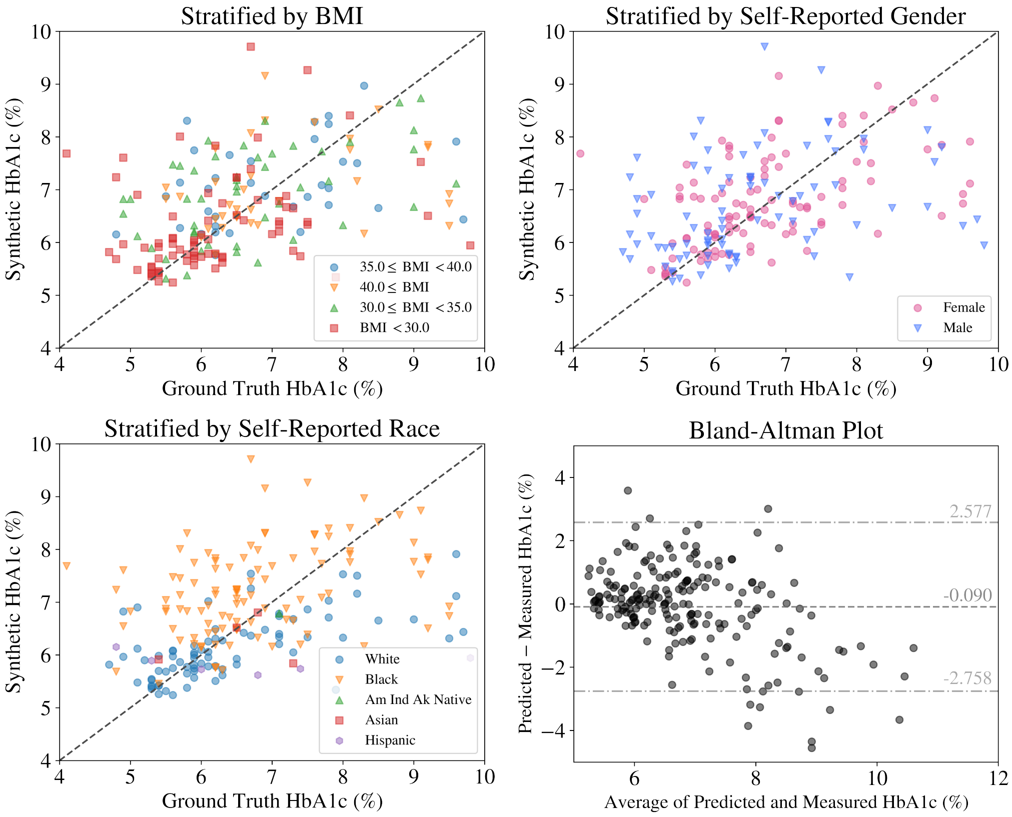

Figure B3: Quantitative analysis of the p-GBDT SynthA1c encoder described in Table 3. (**Top left**) Scatter plot of SynthA1c model outputs vs ground truth HbA1c lab measurements for $N = 208$ test datapoints, stratified by BMI category. (**Top right**) Same plot stratified by self-reported gender. (**Bottom left**) Same plot stratified by self-reported race. (**Bottom right**) Bland-Altman plot comparing SynthA1c and HbA1c values.

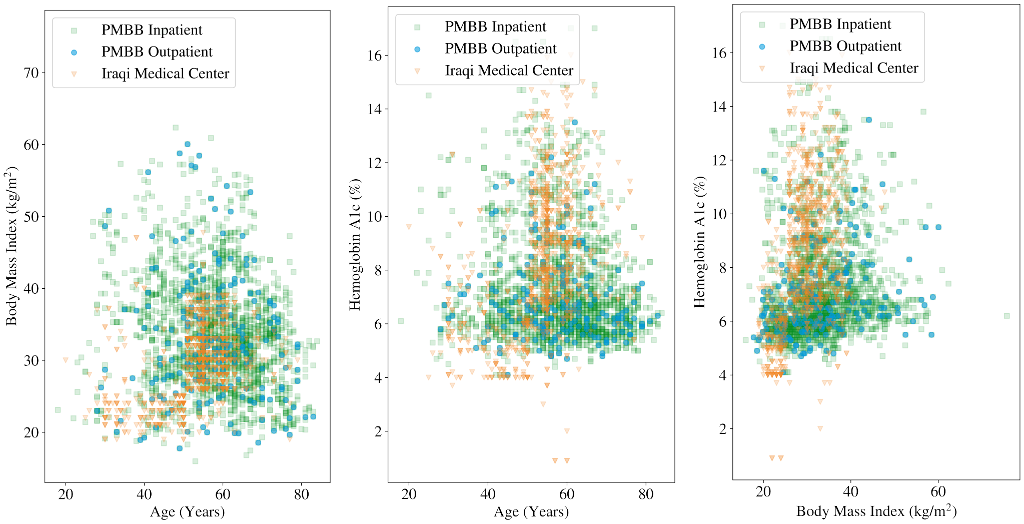

Figure B4: Pair plots of the quantitative features that are available in all three datasets considered in our work: (1) body mass index (BMI), (2) age, and (3) Hemoglobin A1c measurement. The datasets shown here include: (1) the AIBB outpatient test dataset ($N = 208$), (2) the AIBB inpatient dataset ($N = 2066$), and (3) the Iraqi Medical Center dataset ($N = 1000$).