Per-Patient HbA1c Prediction and Biomarker Attribution Using XGBoost and SHAP

Caleb Bucci
University of Texas at Austin
cbucci@utexas.edu

1 Introduction

During the past two decades, diabetes has become an increasingly severe public health crisis in the United States, with the total prevalence of diabetes increasing from 9.7% of adults 20 and older to 14.3% in the 2021-2023 period [16]. There are an estimated 11 million Americans currently living with undiagnosed diabetes, and 115.2 million Americans who are classified as prediabetic [7]. As diabetes prevalence rises, these trends highlight the increasing need for tools that identify risk factors at an individual level. Identifying and predicting risk levels gives individuals the opportunity to make informed lifestyle changes before their condition worsens.

Glycated hemoglobin, or HbA1c, gives an individual's average blood glucose level during the previous two to three months, and it is widely accepted as a reliable marker for the diagnosis of diabetes [27]. A model capable of predicting HbA1c levels from other biomarker data could enable the identification of at-risk individuals before they receive a formal diagnosis. Most existing models approach this problem at a population level [12, 20], which helps identify trends but offers no individual-level assistance.

This project hopes to address that gap by training an XGBoost regression model on NHANES data from 2021-2023 to predict continuous HbA1c levels. After training the model to predict HbA1c levels, SHAP feature attribution will be applied in order to identify the most influential biomarker for each patient. Two models are compared: Model A, which includes fasting glucose, and Model B, which excludes it due to its strong correlation with HbA1c [17]. The two models are to determine whether HbA1c levels can be sufficiently predicted without access to glucose. This work hopes to answer these two questions:

1. Can a supervised model training on NHANES biomarker data accurately predict HbA1c levels at the individual level?
2. Do the resulting SHAP attributions align with or diverge from established clinical guidelines in existing literature?

2 Background

2.1 Diabetes and HbA1c

Diabetes is a chronic disease that occurs when the pancreas does not produce enough insulin or when the body cannot effectively use the insulin it produces, and insulin is a hormone that regulates blood glucose [31]. There are two main types of diabetes: Type 1 and Type 2. Type 1 refers to when the body doesn't make enough insulin; there's no known way to prevent it, and it can develop at any age. Type 1 accounts for approximately 5-10% of all diagnosed cases of diabetes [6]. Type 2 refers to when the body can't use insulin properly; it can also develop at any age, but most cases are preventable. Type 2 accounts for 90-95% of all diagnosed cases [6]. There are many risk factors for type 2 diabetes: being overweight, having a family history, being physically inactive, and being 45 or

Preprint. Under review.

older [6]. While this project does not exclude type 1 patients, the dataset used is predominantly type 2 due to its prevalence. People with diabetes have a higher risk of health complications such as heart attack, stroke, and kidney failure [31]. Early diagnosis is critical to preventing these complications. Diabetes is diagnosed primarily using A1C (also referred to as HbA1c) levels, as seen in Table 1. Diabetes can also be diagnosed using fasting plasma glucose, see Table 2. Fasting plasma glucose requires an 8-hour fast and is a measure of glucose concentration present in an individual's blood at a given point in time, while the HbA1c test requires no fasting and is a marker of the average glucose levels in an individual's blood spread out over a two to three month period [27, 1]. For these reasons, HbA1c is typically the preferred biomarker used for diagnosing diabetes [27, 1].

Table 1: ADA A1C Classification Thresholds [1]

Classification	**A1C Level**
Normal	Below 5.7%
Prediabetes	5.7% – 6.4%
Diabetes	6.5% or above

Table 2: ADA Fasting Plasma Glucose Classification Thresholds [1]

Classification	**Fasting Plasma Glucose Level**
Normal	Below 100 mg/dL
Prediabetes	100 mg/dL – 125 mg/dL
Diabetes	126 mg/dL or above

2.2 NHANES

The National Health and Nutrition Examination Survey (NHANES) is a national survey that measures the health and nutrition of adults and children in the United States, and it is the only national health survey that includes health exams and laboratory tests for participants of all ages [23]. It is conducted by the National Center for Health Statistics (NCHS), a division of the Centers for Disease Control and Prevention (CDC). Examples of ways in which NHANES findings are typically used: determining how common major diseases and behaviors that increase the risk of death, disease, or injury (risk factors) are, assessing nutritional status and how it relates to promoting good health and preventing disease, and developing national estimates for obesity, diabetes, and blood pressure [23]. This project uses the continuous NHANES dataset from 2021-2023 [24], primarily using the examination and laboratory data to retrieve each patient's biomarkers. Additionally, this project does not use the NHANES survey weights, so the results should not be generalized to the US population but rather interpreted at the individual level.

2.3 XGBoost

XGBoost, or Extreme Gradient Boosting, is a scalable tree boosting system used to achieve state of the art results on many machine learning problems [8]. Gradient boosting works by constructing models sequentially where each weak learner minimizes the residual error of the previous one using gradient descent [14, 13]. Gradient boosting of regression trees produces highly robust and interpretable procedures for regression on less than clean data [13], which makes it suitable for the NHANES dataset where missingness is common due to the survey nature of NHANES. Notably, XGBoost handles missing values and NaNs natively during training, which allowed the incomplete data to be handled directly without imputation [8], i.e., replacing missing data with estimated values.

2.4 SHAP

SHAP, or SHapley Additive exPlanations, is a framework for explaining the output of a machine learning model. It assigns each feature an importance value for a prediction, representing how much that particular feature pushed the model output higher or lower compared to the base prediction [21]. SHAP produces these feature importance scores on a per-patient basis rather than a global feature score averaged across all predictions [21]. This makes SHAP well-suited to this project

because it can be used to determine which biomarkers are most strongly associated with predicting HbA1c levels on an individual level. SHAP documentation provides a TreeExplainer algorithm that efficiently computes SHAP values for tree models like XGBoost, and has shown it can be directly applied to NHANES data [26]. One thing to note is that SHAP can show association, but not prove causality [21]; this means that changing the top feature according to SHAP may or may not improve HbA1c levels in this project.

3 Data

The data used in this project comes from the continuous NHANES 2021-2023 cycle, which is split across multiple XPT files throughout the CDC dataset [24]. The data is organized by category: demographics, dietary, examination, laboratory, questionnaire, and limited access data [24]. Since the project was based on biomarkers, only the demographic, examination, and laboratory data were used. There were 34 XPT files in total to be merged. Each row had a unique patient identifier SEQN, which is present across all NHANES files and acts as a primary key [25]. After merging the files and prior to any other data preprocessing, there were 11933 rows (patients) and 425 columns (features) in the dataset. The exact data files used can be seen in data_merge.ipynb. Pandas was used to handle data preprocessing [29].

Demographic features, such as race and gender, were excluded from the model to avoid encoding biases into the model's predictions [34]. The only features saved from the demographics data were the key (SEQN) and the age of patients as that had a direct connection to the likelihood of diabetes in a patient [6]. Survey weights were also dropped from the dataset. These are primarily used for applying findings to a population level [25], and since this is an individual-level prediction model rather than population-level, weights were not applicable.

As a general rule during the data preprocessing, only raw measurement values were kept while the following were dropped:

1. Redundant columns, i.e., columns representing the same measurement in different units, or data representing a ratio/combination of other preexisting values.
2. Comment code columns, typically columns that contained metadata about the raw values themselves, typically derived from the raw values and didn't provide any values for SHAP to work with.
3. All other columns with a high missingness value of $> 15\%$, unless the columns had a direct connection to obesity or diabetes in existing literature.

Values that were coded as refused or unknown (e.g., 7, 77, 777, 9, 99, 999) according to NHANES [25] were recoded to NaN to allow XGBoost to handle them natively. According to NHANES, if 10% or less of data for the main outcome variable (HbA1c) is missing, it is usually acceptable to continue your analysis without any adjustments [25]. To achieve this, all rows that did not have a valid HbA1c score were dropped, resulting in a reduction from 11933 rows to 6686 rows.

There were a few specific exceptions to these general rules. For the examination data, the average of the three systolic and diastolic blood pressure readings were calculated for each and used rather than any one reading to get a more accurate value. BMI, waist circumference, and hip circumference were selected due to connections to obesity in related work [12], while the rest were dropped due to high missingness. In the laboratory data, the BIOPRO_L file, or the standard biochemistry profile dataset, contained readings for triglyceride measurements, glucose measurements, and total cholesterol measurements. These readings also existed in other files (TRIGLY_L, GLU_L, TCHOL_L, respectively), and the documentation specified that these readings should be used over the BIOPRO_L readings [24], so those columns were dropped as well. The columns for Fasting Glucose (LBXGLU), Triglycerides (LBXTLG), Insulin (LBXIN), and LDL-Cholesterol (LBDLDL) all had a missingness value of around 50%. This was because these were in the fasting subsample of patients [24], and due to their high association with diabetes, they were still included despite their missingness being above 15%.

After all data preprocessing, the 11933 x 425 size dataset had shrunk down to a 6686 x 124 size dataset. There were 1125 complete rows, which meant that 1125 patients had laboratory values for every single column in our dataset. In the entire dataset, no column had more than 15% of its values

missing (except for the fasting subsample). The data was then split into two datasets: Glucose-Included (Model A) and Glucose-Free (Model B). Due to the strong correlation between fasting glucose and HbA1c [17], the glucose column was dropped from Model B's dataset to compare the predictive power of the two models and exclude trivial results.

4 Methodology

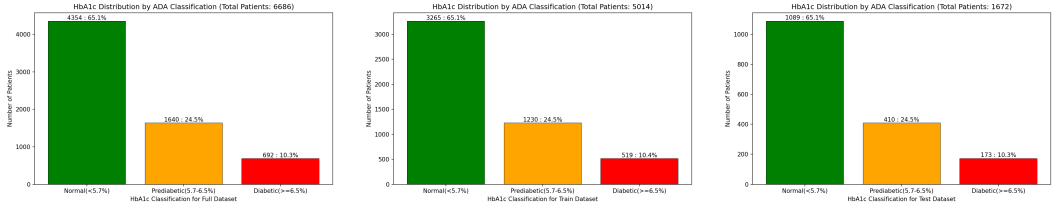
4.1 Model Configuration

Both models were trained using XGBRegressor, which was the Scikit-Learn wrapper interface for XGBoost's regression implementation [33]. The hyperparameters were selected based on ranges found in Khurshid et al., which used Bayesian optimization to optimize XGBoost for enhanced prediction of diabetes [18]. The hyperparameters were set to values within those ranges, as seen in Table 3. The models were trained on a CUDA-enabled gpu, and the random state was set to 33 to enable reproducibility; all other model parameters were left to default.

Table 3: XGBoost Hyperparameters

Hyperparameter	**Value**
n_estimators	500
max_depth	5
learning_rate	0.05
subsample	0.75
colsample_bytree	1.0
gamma	0.2
reg_alpha	1
reg_lambda	1
min_child_weight	5
random_state	33
device	CUDA

Prior to training, each dataset was split into a train and test at a 75/25 split. The full dataset had 10.4% diabetic, 24.5% prediabetic, and 65.1% healthy patients, which is comparable to the American population that has a diagnosed diabetic rate of 10.1% [16]. Stratification was performed on the train and test set using the HbA1c thresholds to ensure each split contained an appropriate portion of normal, prediabetic, and diabetic patients [28]. The exact data splits for each set can be seen in Figure 1.



(a) Full Dataset (b) Train Set (c) Test Set

Figure 1: HbA1c distribution by ADA classification across full, train, and test sets.

4.2 Model Evaluation

Model performance was evaluated using three types of regression metrics: mean absolute error (MAE), root mean squared error (RMSE), and R-squared score. MAE measures the average absolute difference between actual and predicted, giving an easy to interpret average error in the same units as HbA1c [15]. RMSE is the square root of the average of the squared differences between actual and predicted values, which penalizes larger mistakes more heavily than MAE [15]. R-squared represents the proportion of variance in the target variable which provides insight into how well

a model captures underlying data patterns; a value of 1 here would indicate a perfect fit and 0 would indicate the model predicts the mean every time [15]. On top of using these regression metrics, 10-fold cross validation was performed on the full dataset to provide a more robust metric of performance [28, 4].

4.3 SHAP Attribution

After training, SHAP values were calculated with the TreeExplainer algorithm [26] on the entire dataset rather than just the test set, so every patient can receive a SHAP contribution score for each biomarker, representing how much that biomarker increased or decreased the predicted HbA1c [21]. The feature with the highest absolute SHAP value was identified as the most influential biomarker contributing to their predicted HbA1c value. The age column (RIDAGEYR) was excluded from the most influential biomarker recommendations as a patient cannot modify their age. These recommendations were saved to a csv file, where each entry had: SEQN or key, actual HbA1c value, predicted HbA1c value, diabetic class (normal, prediabetic, diabetic), most influential biomarker, value of their most influential biomarker, and the SHAP contribution of their most influential biomarker.

5 Results

5.1 Model Performance

Table 4 shows the test set and cross-validation performance for both models. Model A achieves an R^2 score of 0.6292 and Model B achieves 0.6102, which means that 62.92% of the variance is accounted for in Model A, and 61.02% of the variance is accounted for in Model B. Both models performed rather similarly, with only a 0.019 or 1.9 percentage point absolute difference between the performance of the two models, despite the removal of fasting glucose. Model A's MAE of 0.3721% and Model B's MAE of 0.3897% indicate that predicted HbA1c values differed from their actual HbA1c values by less than 0.4 percentage points, on average. Since the gap between the normal and diabetic thresholds is 0.8 percentage points (from 5.7% to 6.5%), this means that the average prediction of either model would not misclassify a patient as diabetic when they are normal or vice versa [1]. The cross-validation R^2 scores are slightly higher than the test set, which makes sense given that each fold in 10-fold would give a 90/10 split on the training/test fold data, compared to the 75/25 split for the evaluation set. The consistency between the two methods suggests that the models would generalize fairly well to unseen data.

Table 4: Model Performance Metrics

Model	**Mean 10-Fold CV R^2**	**Test RMSE**	**Test MAE**	**Test R^2**
Model A (glucose)	0.6661	0.6603%	0.3721%	0.6292
Model B (glucose-free)	0.6326	0.6770%	0.3897%	0.6102

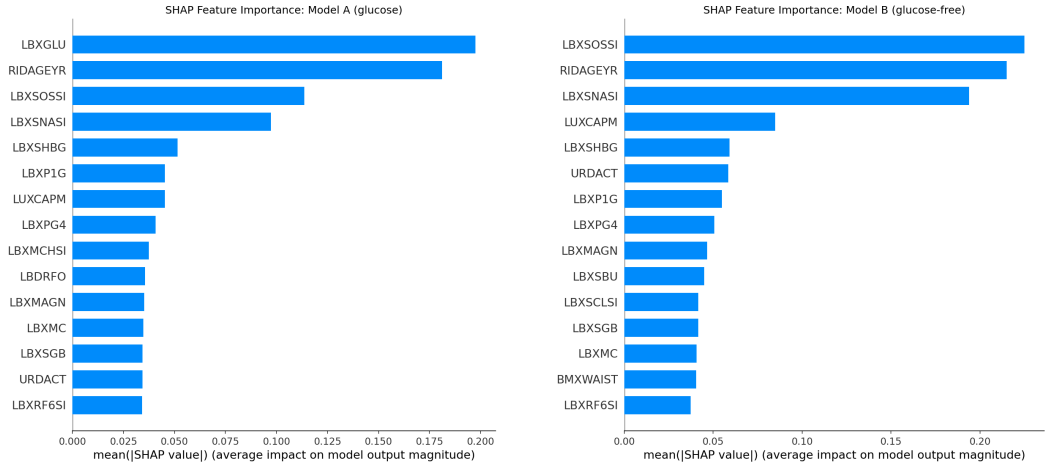
5.2 SHAP Feature Importance

Figure 2 shows the global SHAP feature importance for each model, representing the mean absolute SHAP value for each feature across all patients [21]. The description of each feature is found in the NHANES dataset [24]. In Model A, fasting glucose (LBXGLU) is the strongest feature, followed by age (RIDAGEYR), osmolality (LBXSOSSI), sodium (LBXSNASI), and ending with the sex hormone binding globulin or SHBG (LBXSHBG). In Model B (Glucose-Free), osmolality is the strongest feature (LBXSOSSI), followed by age (RIDAGEYR), sodium (LBXSNASI), median controlled attenuated parameter or CAP (LUXCAPM), and again ending with SHBG (LBXSHBG). The top features for each model can be seen in Table 5.

Figure 3 shows the beeswarm plot of the global SHAP features for each model, where each dot represents a single patient. High feature/biomarker values are indicated by red and low biomarker values are indicated by blue, while positive SHAP values push the prediction above the base prediction and negative SHAP values reduce it. In Model A, high glucose values are associated with positive SHAP contributions, so high glucose pushes the predicted HbA1c higher. In Model B, high median CAP values push the predicted HbA1c higher as well. For both models, low sodium values push the predicted HbA1c value higher, indicating a strong inverse relationship between sodium and

Table 5: Top 5 SHAP Features by Model

Model A (glucose)		Model B (glucose-free)	
Variable	**Description**	**Variable**	**Description**
LBXGLU	Fasting Glucose	LBXSOSSI	Osmolality
RIDAGEYR	Age	RIDAGEYR	Age
LBXSOSSI	Osmolality	LBXSNASI	Sodium
LBXSNASI	Sodium	LUXCAPM	Median CAP
LBXSHBG	SHBG	LBXSHBG	SHBG

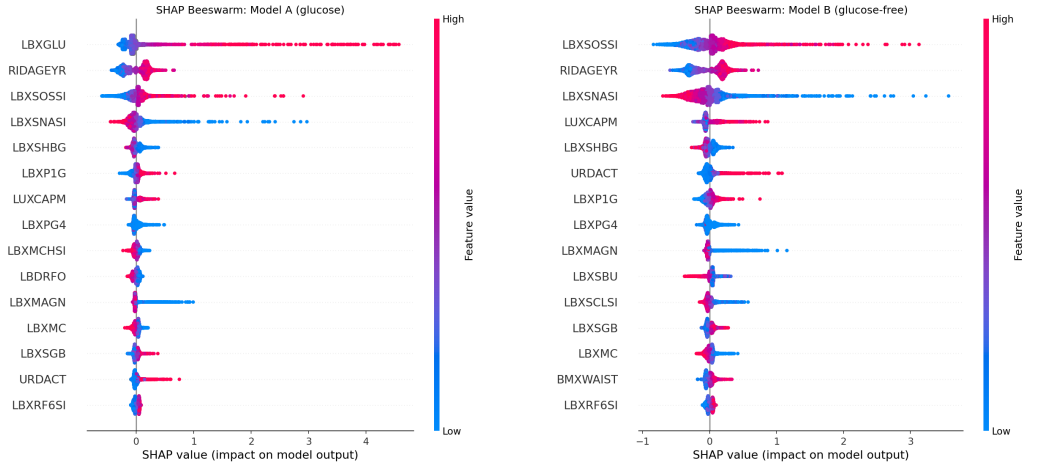


(a) Model A (glucose)

(b) Model B (glucose-free)

Figure 2: SHAP mean absolute feature importance for both models.

HbA1c. Low SHBG values also push the HbA1c value higher which indicates an inverse relationship between SHBG and HbA1c, although the signal isn't as strong. Both high age and osmolality values push the predicted HbA1c value higher.



(a) Model A (glucose)

(b) Model B (glucose-free)

Figure 3: SHAP beeswarm plots for both models (red = high feature value, blue = low).

6 Discussion

6.1 Model Predictive Performance

Both models are moderately capable of predicting HbA1c values, with Model A achieving an R^2 of 0.6292 and Model B achieving 0.6102. A "good" R^2 is difficult to evaluate in clinical research, but an R^2 of $> 15\%$ is generally considered a meaningful value [5], which makes both models substantial. Model A achieved an MAE of 0.3721% and Model B achieved an MAE of 0.3897%. The NHANES laboratory procedure to obtain HbA1c levels in a lab setting uses the HbA1c CAP Total Allowable Error criteria, which is a pass/fail criteria of $\pm 6\%$ [11]. This corresponds to a total absolute allowable criteria of $\pm 0.39\%$ when applied to the diabetes threshold of 6.5% [1]. Both models achieve a tolerance on their average predictions comparable to this threshold. In addition, this benchmark applies to measurements in a controlled lab environment as opposed to predicting HbA1c from a panel of completely different biomarkers, which makes the performance of both models rather notable. The small difference in performance between Model A and Model B suggests that non-glucose related features retain a strong predictive power on their own when it comes to predicting HbA1c.

6.2 SHAP Attribution Analysis

SHAP identified six unique biomarkers across the top five features for both Model A and Model B: Fasting Glucose, Age, Osmolality, Sodium, SHBG, and Median CAP. Both fasting glucose and age have direct connections to diabetes discussed throughout this paper, being 45 or older greatly increases risk of diabetes, which is consistent with the SHAP values that show higher age pushes the predicted HbA1c higher [1, 2, 31]. Hyperglycemia, otherwise known as high blood sugar, affects people who have diabetes [22]; hyperglycemia increases serum osmolality which results in water moving out of the cells and causes a reduction in serum sodium levels [19], which is consistent with SHAP's strong association between high osmolality and increased HbA1c levels. Cheng et al. have also identified an inverse association between serum sodium levels and diabetes in hypertensive patients, which supports SHAP's findings [9]. The typical function of a sex hormone binding globulin is to act as a transport protein for sex steroids, and Wallace et al. have found an inverse connection between low SHBG levels and insulin resistance [30]. Further observational studies [30] have associated a low SHBG concentration with an increased incidence of type 2 diabetes in both men and women, which is consistent with SHAP's findings. Wu et al. ran a cross-sectional study to determine the optimal cutoff value of CAP for predicting type 2 diabetes, and found that participants in the high-CAP group had a significantly greater risk of developing T2DM than those in the low-CAP group, showing it can be used as an auxiliary indicator of type 2 diabetes [32]. The top five global biomarkers for each model that SHAP found to be significant in predicting diabetes have all been found to be supported in existing clinical literature, which suggests that the models have learned clinically meaningful relationships among features.

6.3 Case Studies

To demonstrate the per-patient SHAP attribution system, one patient from each diabetic class was randomly selected from the Model B (glucose-free) recommendation output, and is examined below. Each individual's actual HbA1c, predicted HbA1c, diabetic classification, top contributing SHAP feature, and the corresponding SHAP value are shown in Table 6.

Table 6: Case Study Patient Attributions (using Model B)

Patient	SEQN	Actual	Predicted	Class	Top Feature	SHAP
Patient 1	130378	5.6%	6.04%	Normal	LUXCAPM	+0.270
Patient 2	130380	6.2%	6.07%	Prediabetic	LBXTLG	+0.220
Patient 3	130415	13.5%	13.67%	Diabetic	LBXSOSSI	+2.648

Patient 1 has an actual HbA1c of 5.6% which classifies them as "normal" according to diabetes diagnosis standards [1], and the model predicts them as "prediabetic" with a predicted HbA1c value of 6.04%. This is an absolute difference of 0.44 percentage points, which is slightly outside of

the average MAE loss for Model B (0.3897). This also misclassifies the patient as prediabetic rather than normal, which clearly shows a limitation of the model; patients can be misclassified when close to a diagnostic threshold or when a single feature pushes the prediction too high or too low. The top feature for Patient 1 is LUXCAPM (Median CAP value), which pushed the prediction +0.270 points. Median CAP value is an auxiliary marker for diabetes [32], however CAP also provides an immediate assessment of steatosis [10]. Seeking treatment or a diagnosis could help improve this patient's state.

Patient 2 has an actual HbA1c of 6.2% and a predicted HbA1c value of 6.07%, classifying them squarely in the prediabetic range. This is only an absolute difference of 0.13 percentage points, well within the average MAE loss of 0.3897. The top feature is LBXTLG, or triglycerides. Elevated triglyceride levels are common in patients with type 2 diabetes, and there is evidence that lowering triglycerides may reduce cardiovascular risk in patients [3]. Triglycerides are elevated by poor lifestyle habits such as physical inactivity or high sugar intake [3], making this feature directly actionable.

Patient 3 has an actual HbA1c of 13.5% and a predicted HbA1c of 13.67%, making it the most accurate of the three predictions. Patient 3 is classified in the diabetic range. The top feature is LBXSOSSI, or osmolality, and has a SHAP contribution of +2.648 which is the largest contribution by far from any of the patients. This large contribution is supported by the relationship between osmolality and hyperglycemia [19], and the high value shows the severity of this patient's state.

Although not all top features for each patient were in the top five global SHAP features, each top feature had clinical literature supporting its connection to diabetes. Therefore, the model's per-patient SHAP attributions align with established clinical literature.

6.4 Limitations

Several limitations should be considered before interpreting the results of this project. NHANES survey weights were dropped from the dataset, meaning the results cannot be generalized to the population [25]. Second, no causal relationships can be inferred from the model predictions or the SHAP attributions since SHAP can only show association and the NHANES dataset is cross-sectional [25, 21]. Third, the NHANES 2021-2023 cycle was collected under different procedures due to COVID-19 [24]. Age was excluded from the per-patient recommendations for the top SHAP feature despite it being the second strongest global SHAP feature. Around half of the participants were missing values on key fasting features such as glucose, insulin, LDL, and triglycerides since not all participants took part in the fasting subcategory for NHANES [24]. Finally, the dataset includes patients that have both type 1 and type 2 diabetes, despite the primary motivation of the project being on type 2 patients.

7 Conclusion

This paper trained two XGBoost regression models on NHANES 2021-2023 data to predict HbA1c levels for individual patients. It also used a SHAP attribution system to identify the most influential biomarkers for each patient, as well as globally. Model A, which included fasting glucose, achieved an R^2 of 0.6292 and an MAE of 0.3721%, while Model B, which excluded fasting glucose, achieved an R^2 of 0.6102 and an MAE of 0.3897%. Both models showed meaningful prediction accuracy, showing a comparable average prediction error rate to the CAP total allowable error threshold used in clinical labs [11]. These labs are testing for HbA1c directly while the models are attempting to predict HbA1c values from other biomarker data entirely, making this comparison notable. The small difference in prediction accuracy between the two models shows that non-glucose biomarkers are still powerful predictors for HbA1c. The top SHAP attribution features were fasting glucose, osmolality, sodium, age, median CAP, and SHBG, all of which were found to have associations with diabetes in existing clinical literature. The case studies also showed that the SHAP attribution system found clinically relevant recommendations at the individual level. This supports the claim that these models have learned clinically relevant connections to the top biomarkers. This work only proves association, and does not prove that changing any individual biomarker could directly improve one's HbA1c levels. Future work to improve this project could include distinguishing between type 1 and type 2 diabetes further, applying the NHANES survey weights to generalize to a population level, and improving causality claims to prove more than just association between features and diabetes.

8 Appendix

AI was used as a spell/grammar/tone checker to be in line with the standards of an academic paper, as well as general brainstorming about how to structure the overall paper. It was more directly used to help format LaTEX tables cleanly, and to convert sources into LaTEX bib entries so that they may be used in a references.bib file.

References

[1] American Diabetes Association. Diabetes diagnosis, 2024.

[2] American Diabetes Association Professional Practice Committee for Diabetes. 2. diagnosis and classification of diabetes: Standards of care in diabetes—2026. *Diabetes Care*, 49(Supplement_1):S27–S49, December 2025.

[3] Lars Berglund and John D. Brunzell. Triglycerides: Emerging targets in diabetes care? Review of moderate hypertriglyceridemia in diabetes. *Current Diabetes Reports*, 2019.

[4] Tyler J. Bradshaw, Zachary Huemann, Junjie Hu, and Arman Rahmim. A guide to cross-validation for artificial intelligence in medical imaging. *Radiology: Artificial Intelligence*, 5(4):e220232, 2023.

[5] Felix Camirand Lemyre et al. Determining a meaningful R-squared value in clinical medicine. *Academic Medicine & Surgery*, 2024.

[6] Centers for Disease Control and Prevention. Diabetes: A U.S. report card, 2026.

[7] Centers for Disease Control and Prevention. National diabetes statistics report, 2026.

[8] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. *arXiv preprint arXiv:1603.02754*, 2016.

[9] Qi Cheng, Xiaocong Liu, Anping Cai, Dan Zhou, Yuqing Huang, and Yingqing Feng. Serum sodium level is inversely associated with new-onset diabetes in hypertensive patients. *Journal of Diabetes*, 14(12):831–839, 2022.

[10] Victor de Lédinghen, Julien Vergniol, Juliette Foucher, Wassil Merrouche, and Brigitte le Bail. Non-invasive diagnosis of liver steatosis using controlled attenuation parameter (CAP) and transient elastography. *Journal of Hepatology*, 60(5):1026–1031, 2014.

[11] Diabetes Diagnostic Laboratory, University of Missouri. Laboratory procedure manual: Glycohemoglobin (HbA1c), Tosoh G8 HPLC, 2021.

[12] An Dinh, Stacey Miertschin, Amber Young, and Somya D. Mohanty. A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. *BMC Medical Informatics and Decision Making*, 19:211, 2019.

[13] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5):1189–1232, 2001.

[14] GeeksForGeeks. Ml gradient boosting, 2024.

[15] GeeksForGeeks. Regression metrics, 2024.

[16] Jane A. Gwira, Cheryl D. Fryar, and Qiuping Gu. Prevalence of total, diagnosed, and undiagnosed diabetes in adults: United States, august 2021–august 2023. *NCHS Data Briefs*, (516), November 2024.

[17] Endalkachew Befekadu Ketema and Kelemu Tilahun Kibret. Correlation of fasting and postprandial plasma glucose with HbA1c in assessing glycemic control: systematic review and meta-analysis. *Archives of Public Health*, 73:43, September 2015.

[18] Muhammad Rizwan Khurshid, Sadaf Manzoor, Touseef Sadiq, Lal Hussain, Mohammed Shahbaz Khan, and Ashit Kumar Dutta. Unveiling diabetes onset: Optimized XGBoost with Bayesian optimization for enhanced prediction. *PLOS ONE*, 20(1):e0310218, 2025.

[19] George Liamis, Evangelos Liberopoulos, Fotios Barkas, and Moses Elisaf. Diabetes mellitus and electrolyte disorders. *World Journal of Clinical Cases*, 2(10):488–496, 2014.

[20] Maria Lugner, Aidin Rawshani, Edvin Helleryd, et al. Identifying top ten predictors of type 2 diabetes through machine learning analysis of UK Biobank data. *Scientific Reports*, 14:2102, 2024.

[21] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874*, 2017.

[22] Mayo Clinic. Hyperglycemia in diabetes: Symptoms and causes, 2024.

[23] National Center for Health Statistics. About NHANES, 2024.

[24] National Center for Health Statistics. Nhanes 2021-2023: Continuous datasets, 2024.

[25] National Center for Health Statistics. Nhanes tutorials: Datasets, 2024.

[26] SHAP Community. NHANES I survival model — SHAP documentation, 2024.

[27] Subuhi Imtiyaz Khan Sherwani, Haseeb Ahmad Khan, Aishah Ekhzaimy, Afshan Masood, and Meena Kishore Sakharkar. Significance of HbA1c test in diagnosis and prognosis of diabetic patients. *Biomarker Insights*, 11:95–104, July 2016.

[28] Statology. A complete guide to cross-validation, 2024.

[29] The Pandas Development Team. pandas: Python data analysis library, 2024.

[30] Ian R. Wallace, Michelle C. McKinley, Patrick M. Bell, and Steven J. Hunter. Sex hormone binding globulin and insulin resistance. *Clinical Endocrinology*, 78(3):321–329, 2013.

[31] World Health Organization. Diabetes fact sheet, 2024.

[32] Xingye Wu, Ruibing Qi, Jiaqin Jiang, Quan Lu, Xiaoying Chen, Jiacheng Mo, Jinming Yu, and Zhengming Li. Correlation between the controlled attenuation parameter and the prevalence of T2DM in southern China: A retrospective cross-sectional study. *Diabetes, Metabolic Syndrome and Obesity*, 18:3705–3716, 2025.

[33] XGBoost Developers. XGBoost python API reference, 2024.

[34] Md Rahat Shahriar Zawad and Peter Washington. Evaluating fair feature selection in machine learning for healthcare. *arXiv preprint arXiv:2403.19165*, 2024.