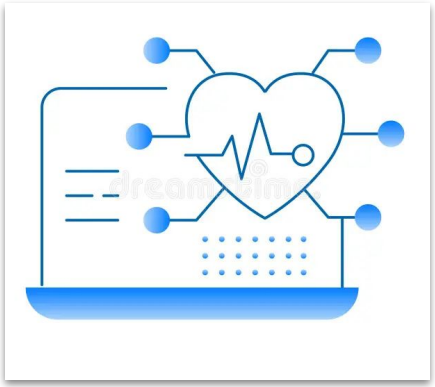
AI Colab Group 1 - Clinical Data Science & Modeling

Predicting β-cell Dysfunction from Kidney Function & Lifestyle Factors in Type 2 Diabetes Using Machine Learning

Yara Yaghi (GMU), Naod Dawit (UMD), Kareem Aly (CVHS), Indira Kuppa (GMU)

AI CoLab, MedStar Georgetown University Hospital



Proposed Research Question

Among U.S. adults with type 2 diabetes, how is kidney function, as measured by estimated glomerular filtration rate (eGFR) and blood urea nitrogen (BUN), associated with β-cell dysfunction, and how do smoking and alcohol behaviors influence this relationship?

Potential Hypothesis

Among U.S. adults with type 2 diabetes, reduced kidney function — indicated by lower eGFR and higher BUN levels — is associated with decreased β-cell function (as measured by HOMA-B). This association is stronger among individuals who currently smoke or consume alcohol more frequently.

Background

- Type 2 Diabetes (T2D) affects over 33 million U.S. adults and contributes to cardiovascular disease, kidney failure, and early mortality (CDC, 2024)
- β-cell dysfunction, a central contributor to T2D progression, impairs insulin secretion and is often under-assessed in risk assessment models (Dludla et al., 2023)
- Kidney function markers (e.g., eGFR, BUN) may serve as early predictors of β-cell function decline
- Lifestyle behaviors (e.g., smoking, alcohol use) are associated with worsened metabolic function, potentially amplifying T2D progression (Akhuemonkhan & Lazo, 2017)
- Most clinical risk prediction models for Type 2 Diabetes focus on lifestyle factors, age, BMI, glucose, and family history, while excluding β-cell function markers.
- Existing models often overlook the combined impact of biological and behavioral risk factors on β-cell health.

Why AI/ML

- Machine learning models capture non-linear, interactive relationships, improving prediction of β-cell dysfunction.
- Incorporates a broad spectrum of features (clinical + behavioral) to enhance risk stratification
- Enables early identification of at-risk patients for personalized intervention

Our work:

- Focuses on HOMA-B as a measure of β-cell function, offering greater precision than basic glucose thresholds
- Combines renal biomarkers and lifestyle behaviors data using a large, nationally representative dataset (NHANES)
- Emphasizes model interpretability to clearly link input variables with outcomes in clinically meaningful ways

Objectives

1. To evaluate the association between kidney function biomarkers (eGFR and BUN) and β-cell dysfunction (measured by HOMA-B) among U.S. adults with type 2 diabetes using machine learning models.

2. To assess how smoking and alcohol use influence the relationship between kidney function and β-cell dysfunction

3. To develop and validate predictive models capable of accurately identifying patients at risk of β-cell dysfunction

4. To inform early intervention strategies and contribute to precision medicine for diabetic patients

DATA SOURCE

- Dataset: National Health and Nutrition Examination Survey (NHANES)
- Years Covered: 1999–2020 (Multiple 2-year cycles combined)
- Population: U.S. adults aged 30 and above
- Source Website: https://wwwn.cdc.gov/nchs/nhanes/

NHANES Data Modules Used (2007–2017):

- Demographics Module
- Laboratory Module
  - Fasting Glucose & Insulin (LBXGLU, LBXIN)
  - Kidney Function Biomarkers: BUN & Creatinine (LBXSBU, LBXSCR)
  - Urine Albumin-Creatinine Ratio & Components (URDACT, URXUMA, URXUCR)
- Diabetes Questionnaire Module
- Smoking Questionnaire Module
- Alcohol Use Questionnaire Module

Inclusion/Exclusion Criteria

Included:

- Participants from NHANES cycles 1999-2020
- Age ≥ 30
  - To minimize inclusion of early-onset or Type 1 diabetes
- Has Type 2 diabetes (self-reported)
- Available fasting glucose and fasting insulin values

Excluded:

- Missing key health variables
- Missing demographic & behavioral variables
- eGFR < 30 mL/min/1.73m^2
  - Indicates severe chronic kidney disease (Stage 4+)
- Participants without diabetes (self-reported)

Methodology

Phase 1: Data Merging & Filtering

- Merge all years' dataset
- Select only features required for analysis
- Standardized all the features from all datasets and created one master dataframe

- Selected only patients with diabetes
- Flagged likely Type 1 Diabetes based on:
  - Diagnosed before age 30
  - Started insulin within one year
  - Not currently on diabetes pills or missing pill info

- Kept only Type 2 diabetes patients

Phase 2: Data Clean & Engineer

- Handled missing values
  - Used mean, median, and logical imputation
- Handling outliers
- Dropping features with strong correlation (multicollinearity)

- Feature Engineered:
  - HOMA_B
  - ACR (urine albumin-to-creatinine ratio)
  - Excluded participants with eGFR < 30 mL/min/1.73m^2

- Balanced features (worse than 80:20)
- Took the log & scaled for skewed features & large ranges

Phase 3: Modeling & Evaluation

- Split data into train/test
- Train models (Logistic Regression, XGBoost, RandomForest) on the training set

- Assess model performance using R^2 and MSE

Methodology

BEFORE

	Feature Name	Data Type	Missing Values	Description
0	SEQN	float64	0	Respondent sequence number (unique ID for each...
1	LBXGLU	float64	1489	Fasting glucose (mg/dL)
2	LBXIN	float64	2015	Fasting insulin (µU/mL)
3	LBXSBU	float64	1801	Blood urea nitrogen (BUN) (mg/dL), marker of k...
4	LBXSCR	float64	1800	Serum creatinine (mg/dL), used to assess kidne...
5	LBXSATSI	float64	1826	Serum sodium concentration (mmol/L)
6	SMQ020	float64	140	Ever smoked at least 100 cigarettes in life (1...
7	SMQ040	float64	15294	Current smoking status (1 = Every day, 2 = Som...
8	EverDrank	float64	17061	No description available
9	DrinkFrequency	float64	6055	No description available
10	AvgDrinksPerDay	float64	11018	No description available
11	RIDAGEYR	float64	0	Age in years at time of screening
12	RIAGENDR	float64	0	Gender (1 = Male, 2 = Female)
13	DMDEDUC2	float64	628	Education level (1 = Less than 9th grade to 5 ...
14	INDFMPIR	float64	2668	Ratio of family income to poverty level (highe...
15	DIQ010	float64	0	Doctor told you have diabetes (1 = Yes, 2 = No)
16	DID040	float64	24323	Age when first told you had diabetes
17	DIQ050	float64	4369	Currently taking insulin (1 = Yes, 2 = No)
18	DIQ070	float64	22596	Currently taking pills to lower blood sugar (1...
19	SurveyCycle	object	0	NHANES survey cycle years
20	URDACT	float64	641	Urine albumin-to-creatinine ratio (mg/g), mark...

AFTER

	Feature Name	Data Type	Missing Values	Description
0	BUN_scaled	float64	0	Standardized blood urea nitrogen level (marker...
1	EverSmoked100	float64	0	Ever smoked at least 100 cigarettes in life (1...
2	Smoker	int64	0	Current smoker flag (1 = Yes, 0 = No)
3	eGFR_scaled	float64	0	Standardized estimated glomerular filtration r...
4	ACR_log	float64	0	Log-transformed urine albumin-to-creatinine ra...
5	EverDrank_Label_Drinks Alcohol	int64	0	Label indicating whether participant drinks al...
6	AvgDrinksPerDay	float64	0	Average number of alcoholic drinks per day ove...
7	Gender	float64	0	Gender (1 = Male, 2 = Female)
8	Age_scaled	float64	0	Standardized age of participant
9	Education	float64	0	Education level (1 = Less than 9th grade to 5 ...
10	IncomeToPovertyRatio	float64	0	Ratio of family income to poverty level
11	HOMA_B_scaled	float64	0	Scaled Homeostatic Model Assessment of Beta-ce...

Modeling/Technical Approaches

ML Algorithms Used:

- **Random Forest** (baseline ensemble)
- **XGBoost** (gradient boosting)
- **Bagging** (ensemble blending)
- **LightGBM** (fast & scalable)

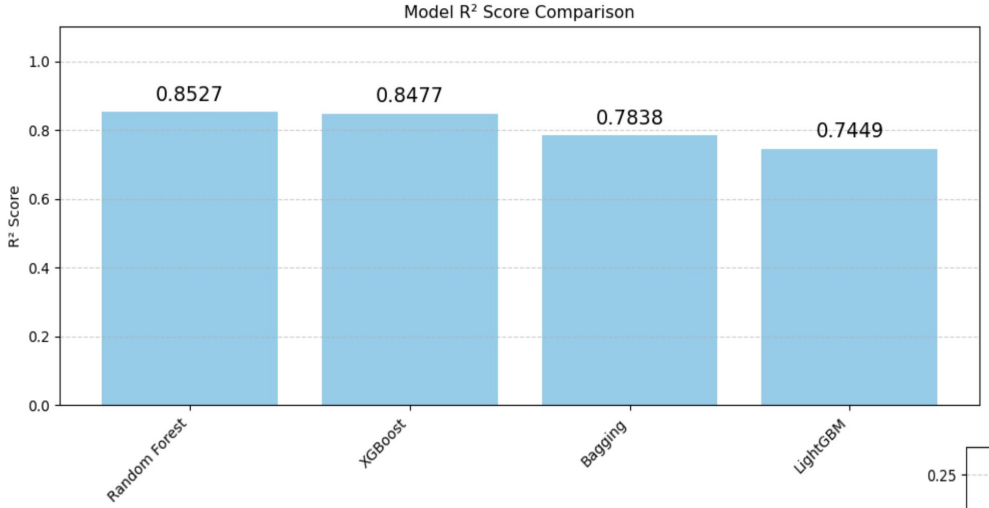
Tools & Libraries:

- Python (Jupyter Notebook)
- scikit-learn, xgboost, lightgbm, matplotlib, pandas, numpy
- GridSearchCV for hyperparameter tuning

Performance Metrics:

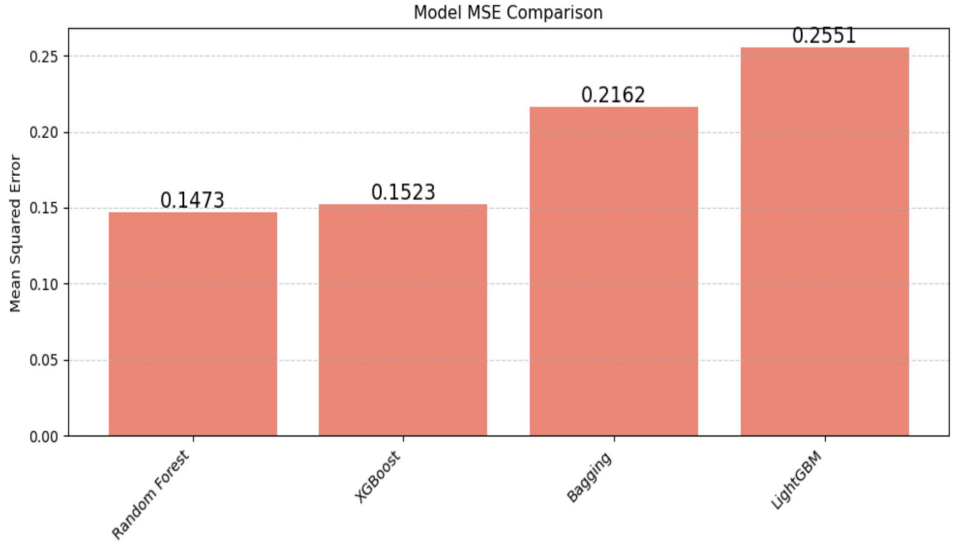
- **R^2 Score** – goodness of fit
- **MSE** – average squared error
- **Feature importance plots** for interpretability

Finding/Results



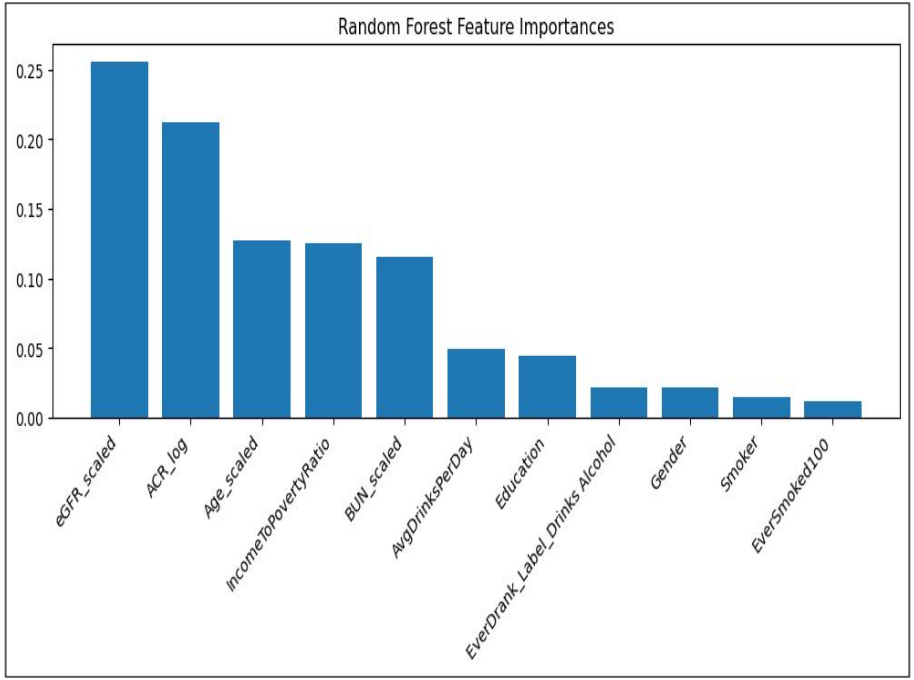
Model	R^2
Random Forest	**0.8527**
XGBoost	0.8477
Bagging	0.7838
LightGBM	0.7499

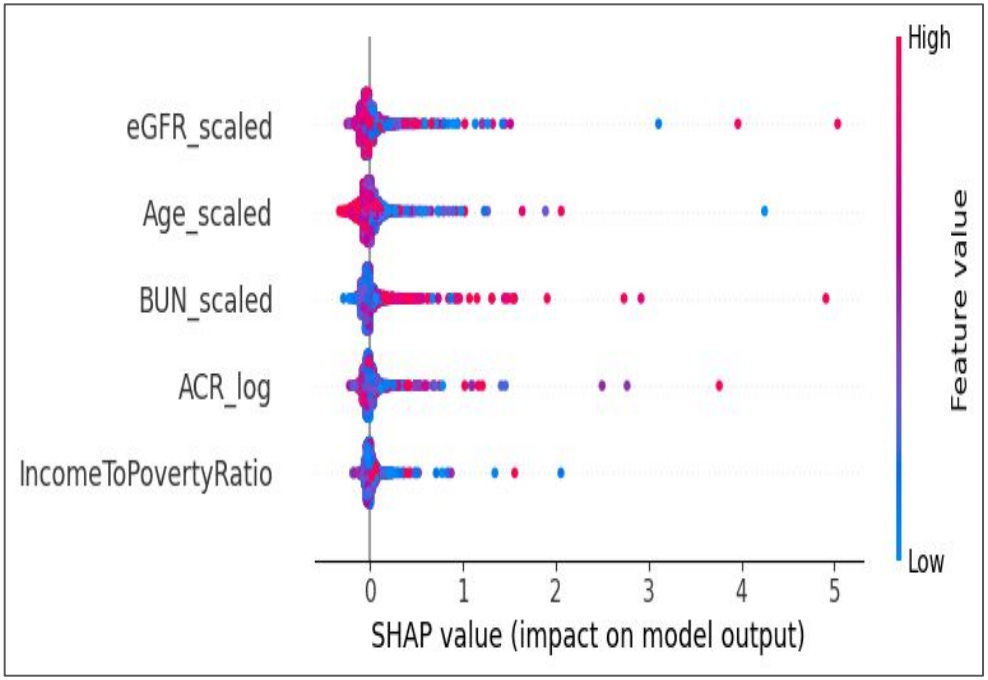
Model	MSE
Random Forest	**0.1473**
XGBoost	0.1523
Bagging	0.2162
LightGBM	0.2551



Results

```
Random Forest MSE: 0.1469
Random Forest R²: 0.8531
```





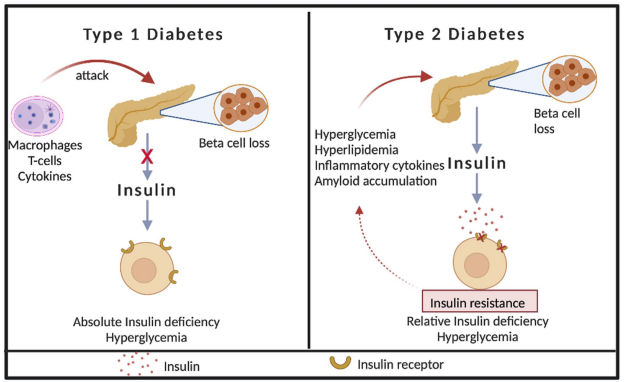
Impacting/Healthcare Lens

Our project mainly advances both population health and personalized approaches in diabetes care, paving the way for future innovation in terms of Type 2 Diabetes and Kidney Function.

- Our project investigates how **kidney function (eGFR, BUN)** and **lifestyle factors (smoking, alcohol use)** relate to **β-cell dysfunction** in adults with **Type 2 Diabetes.**
- Working on Predictive analytics enables early identification of at-risk patients, supporting personalized treatment and improved glycemic control.
- Aids as a foundation for Precision Medicine- Integrates behavioral and biological markers to better tailor diabetes care according to individual risk profiles.

<u>Biomedical Relevance</u>

- Endocrine disruptors (EDCs) interfere with pancreatic β-cell function by causing mitochondrial and cellular stress, leading to impaired insulin secretion and reduced β-cell mass, which promotes Type 2 Diabetes development.



Challenges & Lessons Learned

- NHANES dataset lacked key clinical biomarkers (e.g., HbA1c, C-peptide, medication data), constraining model depth.
- Merging multiple NHANES cycles and addressing missing data demanded extensive cleaning, standardization, and scaling.
- Data were cross-sectional, not longitudinal, limiting the ability to track features incline or decline over time.

Lessons Learned

- Careful selection, scaling, and transformation of features are critical for optimal model performance and interpretability.
- Thorough data cleaning and preparation are foundational to project success.
- Exploratory data analysis (EDA) is essential for uncovering patterns and informing feature selection.
- Model evaluation metrics (AUC, MSE, feature importance) offer valuable insights into strengths and limitations.
- Learning of the each member's strengths and weaknesses and being able to divide tasks which allowed us to complete this project.

Future Considerations

- Acquire and analyze **real-time patient data** to assess β-cell function more dynamically, enabling comparison with previous cross-sectional findings.
- Select and incorporate **better features**—including additional clinical, laboratory, and behavioral biomarkers—to strengthen model accuracy and clinical relevance.
- Apply **advanced machine learning (ML)** and artificial intelligence (AI) techniques (e.g., ensemble models, deep learning, hybrid approaches) to capture complex relationships and non-linear patterns in the data.
- Validate models using external, prospective data and carry out rigorous evaluation (e.g., AUC, feature importance) to ensure clinical utility.

Machine Learning-Based Prediction of Type 2 Diabetes from Kidney Function and β-cell Dysfunction

Yara Yaghi, Naod Dawit, Kareem Aly, Indira Kuppa

MedStar Health

INTRODUCTION

Type 2 Diabetes Mellitus (T2DM) is a progressive endocrine disorder marked by insulin resistance and β-cell failure. Kidney function, commonly assessed via estimated glomerular filtration rate (eGFR) and blood urea nitrogen (BUN), plays a critical yet underexplored role in β-cell physiology. Emerging research suggests that chronic kidney impairment may exacerbate β-cell dysfunction. Smoking and alcohol use independently contribute to both renal injury and β-cell cytotoxicity, potentially acting as modifiers in the relationship between nephropathy and islet cell function. Gaining insight into these interrelated mechanisms is essential for stratifying endocrine risk, refining therapeutic targets, and advancing precision diabetes care.

OBJECTIVES

1. To evaluate the association between kidney function markers (eGFR and BUN) and β-cell function (HOMA-B) among U.S. adults with type 2 diabetes using regression modeling techniques.
2. To assess the modifying effect of smoking status and alcohol consumption—on the relationship between kidney function and HOMA-B.

METHODS AND MATERIALS

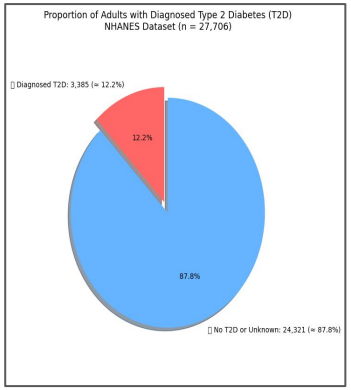
Data Source: National Health and Nutrition Examination Survey (NHANES) (1999–2020)

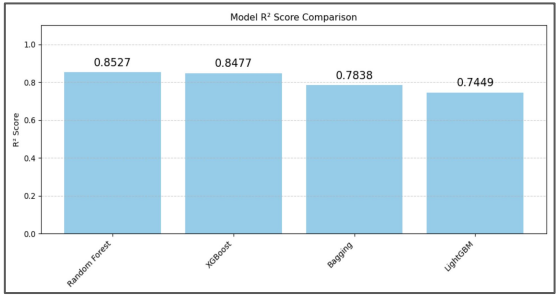
PHASE 1 Data Merging & Filtering-Merged multi-year NHANES data and filtered for Type 2 diabetes cases

PHASE 2 Data Cleaning & Feature Engineering-Cleaned data, handled missing values, and engineered features like HOMA-B and ACR

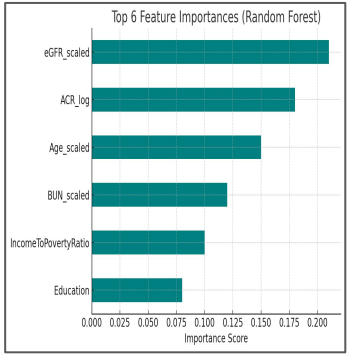
PHASE 3 Modeling & Evaluation- Trained and evaluated predictive models using R^2 and MSE metrics

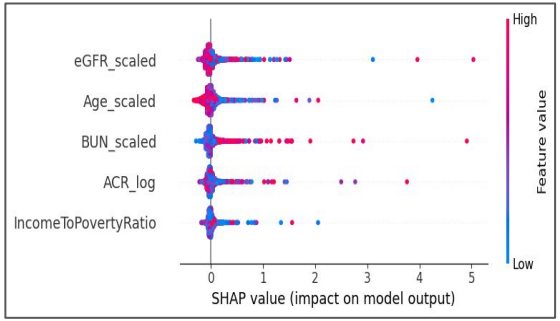
RESULTS





Random Forest	R^2 Score- 0.8527	MSE- 0.1473
XGBoost	R^2 Score-0.8477	MSE- 0.1523





CONCLUSION

1.By merging NHANES multi-year cycles (1999–2020), cleaning and engineering clinical and behavioral features, and applying advanced ML models, we successfully built a predictive framework for β-cell function analysis in T2DM patients.
2..Kidney function markers (eGFR, BUN, ACR) were among the strongest predictors of β-cell function (HOMA-B) in U.S. adults with type 2 diabetes.
2. Random Forest achieved the best predictive performance (R^2 = 0.85, MSE = 0.15), outperforming other machine learning models like XGBoost, Bagging, and LightGBM.
3. The model demonstrated that combining clinical (renal biomarkers) and socio-behavioral factors can provide valuable insights into β-cell dysfunction risk.
4. By using ensemble models and interpretability tools like SHAP, we not only improved prediction accuracy but also gained insight into feature contributions.
5. This work highlights the potential for ML-based risk stratification models to support early intervention strategies in diabetes care. Future work should explore validation on external cohorts and the integration of longitudinal data.

LIMITATIONS AND FUTURE RECOMMENDATIONS

1. NHANES lacked key clinical biomarkers (e.g., HbA1c, C-peptide, medication data), limiting model depth.
2. Combining multiple NHANES cycles required extensive cleaning, standardization, and handling missing data.
3. Cross-sectional data limited tracking of β-cell function changes over time.
4. Acquire and analyze real-time patient data for dynamic β-cell function assessment and comparison.
5. Enhance models with richer clinical, laboratory, and behavioral features to improve accuracy and relevance.
6. Employ advanced ML/AI techniques (ensemble, deep learning, hybrid models) to capture complex patterns.
7. Validate models on external, prospective datasets with rigorous metrics to confirm clinical utility.

Acknowledgements

- Dr. Nawar Shara, PhD - Founding Co-Director, AI CoLab, MedStar Health Research Institute
- Omar Aljawfi - Informatics Analyst & Data Scientist, MedStar Health Research Institute
- Mentors: Raed Darwish, Diyang Lyu, Zannatul Ferdous, William Mea
- Maryam Solimany - Program Coordinator, AI CoLab, MedStar Health Research Institute

References

1. Akhuemonkhan, E. & Lazo, M. (2017). Association between family history of diabetes and cardiovascular disease and lifestyle risk factors in the U.S. population (NHANES 2009–2012). *Preventive Medicine, 96*, 129–134. https://doi.org/10.1016/j.ypmed.2016.12.015
2. CDC. (2024, May 15). About Type 2 Diabetes. Diabetes. https://www.cdc.gov/diabetes/about/about-type-2-diabetes.html
3. Dludla, P. V., Mabhida, S. E., Ziqubu, K., Nkambule, B. B., Mazibuko-Mbeje, S. E., Hanser, S., Basson, A. K., Pheiffer, C., & Kengne, A. P. (2023). Pancreatic β-cell dysfunction in type 2 diabetes: Implications of inflammation and oxidative stress. *World Journal of Diabetes, 14*(3), 130–146. https://doi.org/10.4239/wjd.v14.i3.130

Acknowledgements

We would like to express our sincere gratitude to the individuals and organizations who contributed to the success of this project.

- Dr. Nawar Shara, PhD - Founding Co-Director, AI CoLab, MedStar Health Research Institute
- Omar Aljawfi - Informatics Analyst & Data Scientist, MedStar Health Research Institute
- Mentors: Raed Darwish, Diyang Lyu, Zannatul Ferdous, William Mea
- Maryam Solimany - Program Coordinator, AI CoLab, MedStar Health Research Institute

Data Source

We acknowledge the use of publicly available data from the National Health and Nutrition Examination Survey (NHANES), administered by the CDC, which provided the foundation for our analysis.

References

Akhuemonkhan, E. & Lazo, M. (2017). Association between family history of diabetes and cardiovascular disease and lifestyle risk factors in the U.S. population (NHANES 2009–2012). *Preventive Medicine, 96*, 129–134. https://doi.org/10.1016/j.ypmed.2016.12.015

CDC. (2024, May 15). About Type 2 Diabetes. Diabetes. https://www.cdc.gov/diabetes/about/about-type-2-diabetes.html

Dludla, P. V., Mabhida, S. E., Ziqubu, K., Nkambule, B. B., Mazibuko-Mbeje, S. E., Hanser, S., Basson, A. K., Pheiffer, C., & Kengne, A. P. (2023). Pancreatic β-cell dysfunction in type 2 diabetes: Implications of inflammation and oxidative stress. *World Journal of Diabetes, 14*(3), 130–146. https://doi.org/10.4239/wjd.v14.i3.130