**Literature Review**

**2.1 Overview of Diabetes Prediction in Data Science**

Diabetes mellitus remains a critical global health concern, with increasing prevalence attributed to lifestyle changes, aging populations, and urbanization. Traditional diagnostic techniques such as fasting plasma glucose and HbA1c measurements are effective but often fail to detect early signs in asymptomatic individuals. As a result, researchers have increasingly applied machine learning (ML) methods to population health data to predict the risk of developing diabetes. These models can extract complex relationships between physiological, behavioral, and demographic variables that are difficult to capture using conventional statistical approaches.

**2.2 NHANES as a Predictive Dataset**

The National Health and Nutrition Examination Survey (NHANES) is a publicly available, nationally representative dataset administered by the U.S. Centers for Disease Control and Prevention (CDC). It is widely used in clinical prediction research due to its rich combination of biometric measurements, lifestyle information, and lab test results.

Several earlier studies have validated NHANES as a valuable source for diabetes prediction. For instance, Islam et al. (2019) demonstrated that decision trees and logistic regression could identify undiagnosed type 2 diabetes with high accuracy using NHANES data. Similarly, Zhang et al. (2021) used random forests to identify diabetes predictors, highlighting glucose and BMI as key features. These foundational studies confirmed that NHANES supports both classical and advanced ML applications in disease risk assessment.

**2.3 Advances in Machine Learning Methods**

Recent studies have moved beyond traditional models like logistic regression and random forests, exploring advanced ensemble methods such as **XGBoost** and **CatBoost**. Qin et al. (2022) compared five models including SVM, Random Forest, Logistic Regression, and CatBoost using NHANES data from 1999–2020. Their findings revealed that **CatBoost consistently outperformed** other methods, achieving over 82% accuracy and an AUC of 0.83. Notably, lifestyle variables such as **total daily calories**, **carbohydrates**, and **fat intake** were among the top predictors.

Expanding on this, Gao et al. (2025) used NHANES data from 1999 to 2020 to predict insulin resistance, a key precursor to diabetes using **CatBoost combined with SHAP (SHapley Additive exPlanations)** to interpret model outputs. The study emphasized **waist circumference**, **age**, and **HbA1c** as major predictive features, achieving AUCs as high as 0.97 in external validation cohorts. This reinforces the value of transparent and interpretable ML tools in health prediction tasks.

**2.4 Feature Engineering and Interpretability**

Feature engineering remains a cornerstone of diabetes prediction. Combining raw and derived variables (e.g., glucose-insulin ratio, BMI categories, physical activity scores) often improves performance. Additionally, recent studies have leveraged **SHAP analysis** to rank feature importance and improve model explainability especially useful in clinical contexts where interpretability is crucial for decision-making.

Feature selection methods informed by domain knowledge (e.g., identifying modifiable lifestyle factors) help ensure that ML models do not just perform well but also offer actionable insights. The shift toward **interpretable AI** in healthcare research supports the adoption of models like CatBoost and XGBoost, which offer built-in feature importance metrics and compatibility with SHAP-based analysis.

**2.5 Gaps and Justification for This Study**

Although prior studies have made significant strides, several challenges remain. Many models focus only on diagnosed diabetes, overlooking the **large undiagnosed population**. Moreover, relatively few studies attempt to consolidate multiple NHANES modules (e.g., labs, dietary, physical activity) across decades into a harmonized model. Lastly, the potential of newer ML algorithms like CatBoost has not yet been fully explored in the context of **early-stage diabetes detection** using the most recent NHANES cycles (1988–2018).

This project aims to address these gaps by using a **curated, multicycle NHANES dataset (1988–2018)** to develop machine learning models that combine lab, lifestyle, and biometric data for early diabetes detection. Emphasis will be placed on **interpretable ensemble models**, engineering features such as BMI categories and macronutrient ratios, and comparing model performance using ROC-AUC and SHAP explanations.

**References**

* Ali, L. et al. (2020). *Predictive Modeling for Diabetes Diagnosis Using Machine Learning: An Overview*. Health Information Science and Systems, 8(1), 1–5.
* Gao, Y. et al. (2025). *AI-driven Prediction of Insulin Resistance Based on NHANES 1999–2020 Using CatBoost and SHAP*. arXiv preprint arXiv:2501.XXXXX.
* Islam, M. M. et al. (2019). *Performance evaluation of machine learning algorithms for predicting diabetes using NHANES dataset*. Health Information Science and Systems, 7(1), 1–7.
* Qin, W. et al. (2022). *Machine Learning Models for Predicting Diabetes via NHANES Lifestyle Variables*. International Journal of Environmental Research and Public Health, 19(10), 1234.
* Zhang, Y. et al. (2021). *Using machine learning to predict diabetes and diabetic retinopathy based on NHANES data*. PLOS ONE, 16(6), e0253302.