In the United States and globally, Type 2 Diabetes presents a public health challenge. Early and accurate prediction of T2D can facilitate timely interventions, thereby mitigating the risk of long-term complications. Traditional statistical methods, while valuable, often fall short of capturing the intricate relationships among various T2D risk factors. Machine learning (ML) models, with their ability to handle large, multidimensional datasets, offer promising alternatives. This study explores the application of ML in T2D prediction, leveraging a broad spectrum of predictors from lifestyle habits and socioeconomic backgrounds to environmental exposures, aiming to identify the model that best predicts T2D risk.

The study analyzed data comprising 129,024 individuals, including 21,303 with type 2 diabetes, from the Center for Disease Control, 2014 Behavioral Risk Factor Surveillance System, annotated with variables including BMI, age, exercise frequency, smoking status, sleep duration, socioeconomic status (SES), healthcare access, and geographical factors. The data was split into 80% training and 20% testing sets. Five Machine Learning models (Logistic Regression, Gaussian Naive Bayes, Random Forest, Gradient Boosting, and Decision Trees) were trained, and their performances were compared based on accuracy, sensitivity, specificity, and the Area Under Curve (AUC). Feature importance analysis was conducted to identify the most predictive variables.

The predictive models used achieved a high area under the curve (AUC ranging from 0.69 to 0.81). However, the Gradient Boosting Model (GBM) outperformed others with an accuracy of 0.86, specificity of 0.99, and AUC of 0.81. The Gaussian Naive Bayes model presented a balanced sensitivity-specificity trade-off (Accuracy: 0.80, Sensitivity: 0.43, Specificity: 0.87, AUC: 0.75) but fell short in overall accuracy and AUC compared to Gradient Boosting. This study highlighted that people who are unable to work (Coef = 0.5649), or who had a Health Care Coverage with the Alaska Native/Indian Health Service (Coef = 0.6096), or who have not had a medical checkup in the last 5 years or more (Coef = -1.0077) have higher risk for type 2 diabetes.

The gradient boosting model showed the best model performance with the highest AUC value; however, the naive Bayes model is preferred for initial screening for type 2 diabetes because it had the highest sensitivity and, therefore, detection rate. The superior performance of GBM can be attributed to its ability to handle complex interactions among a range of risk factors, from biomedical to socioeconomic and environmental. Notably, the study confirms previously reported risk factors like BMI, Age, gender, etc. by Gary Collins et al, it also identifies employment status, healthcare coverage type, and frequency of medical checkups as 3 new potential risk factors related to T2D, highlighting the potential for machine learning to uncover nuanced insights into disease prediction. By integrating these broader determinants of health, machine learning models can offer a more comprehensive tool for early disease detection, thus highlighting the critical role of machine learning in advancing personalized medicine, health informatics and public health strategies.