Fang M et al. (2024). INTERPRETABLE MACHINE LEARNING ANALYSIS OF DIET, STRESS, AND THEIR IMPACT ON DIABETES RISK: IMPLICATIONS FOR METABOLIC HEALTH AND PHYSICAL PERFORMANCE. Revista Internacional de Medicina y Ciencias de la Actividad Física y el Deporte vol. 24 (97) pp. 512-531.
DOI: https://doi.org/10.15366/rimcafd2024.97.034

ORIGINAL

INTERPRETABLE MACHINE LEARNING ANALYSIS OF DIET, STRESS, AND THEIR IMPACT ON DIABETES RISK: IMPLICATIONS FOR METABOLIC HEALTH AND PHYSICAL PERFORMANCE

Qun Wan[1], Qian Chen[2,5*], Min Fang[3,*], Xiaozhen Sun[2], Zihang Zhao[4]

[1]Department of Clinical Nutrition, Shenzhen University General Hospital, Guangdong, China.
[2]Computer Science and Technology, Harbin Institute of Technology, Shenzhen, Shenzhen, 518000, Guangdong, China.
[3] Education Center of Experiments and Innovations, Harbin Institute of Technology, Shenzhen, Shenzhen, Guangdong, China.
[4]Computer Science and Technology, University of Hong Kong, Hong Kong, China.
[5]Electronic Commerce, Shenzhen Campus, Jinan University, Guangdong, China.
E-mail: qianchen@stu.hit.edu.cn; fangmin@hit.edu.cn

Recibido 26 de diciembre de 2023 **Received** December 26, 2023
Aceptado 26 de julio de 2024 **Accepted** July 26, 2024

ABSTRACT

Background: Diabetes is a major metabolic disorder that not only affects overall health but also significantly influences physical activity levels, athletic performance, and exercise recovery. Early and accurate detection of diabetes is crucial for preventing complications, optimizing metabolic function, and enhancing participation in physical activity and sports. This study explores the impact of dietary habits and stress factors on diabetes risk using interpretable machine learning models, with a focus on their implications for sports science, rehabilitation, and metabolic health management. **Methods:** Machine learning models were developed using dietary intake data combined with positive and negative stress indicators to enhance predictive accuracy for diabetes detection. Comparative analyses were conducted to evaluate the relative impact of diet and stress on diabetes risk, with an emphasis on metabolic efficiency, energy regulation, and physical endurance. Random Forest and other interpretable machine learning approaches were applied to ensure transparency in the prediction process, enabling clinicians, sports scientists, and health practitioners to derive actionable insights from the results. **Results:** The inclusion of stress-related features significantly improved model accuracy and generalizability, highlighting the interplay between psychological stress, metabolic function, and physical performance. Contrary to traditional

assumptions, positive stress exhibited a stronger influence on diabetes risk than negative stress, suggesting that psychological resilience and adaptive stress responses play a crucial role in metabolic adaptation and physical health. Additionally, dietary factors, particularly carbohydrate intake, emerged as the most critical determinant of diabetes risk, reinforcing the importance of nutritional regulation in sports performance and metabolic optimization. The proposed machine learning model achieved an accuracy exceeding 99%, demonstrating its potential as a reliable tool for early diabetes detection, personalized intervention, and sports health management. **Conclusions:** This study provides valuable insights into the role of diet and stress in diabetes risk and their implications for physical activity, sports participation, and athletic performance. The findings highlight the need for integrated lifestyle interventions that combine nutritional optimization, stress management, and structured exercise programs to enhance metabolic resilience and athletic endurance. By leveraging interpretable machine learning, healthcare professionals and sports scientists can develop personalized strategies for diabetes prevention, physical conditioning, and performance enhancement. Future research should explore the long-term impact of diet-stress interactions on sports performance and recovery in diabetic and prediabetic populations.

KEYWORDS: Diabetes Detection; Interpretable Machine Learning; Dietary Structure; Life Stress

1. INTRODUCTION

Diabetes mellitus (DM) is a global metabolic disorder that significantly impacts physical activity, exercise performance, and overall health. Characterized by insulin resistance, impaired glucose metabolism, and chronic hyperglycemia, diabetes is associated with severe complications such as cardiovascular disease, neuropathy, and musculoskeletal dysfunction. Beyond its clinical implications, diabetes affects an individual's ability to engage in regular physical activity, reducing exercise tolerance, aerobic capacity, and muscle recovery (Rajkomar et al., 2019; Ribeiro et al., 2016). These limitations pose a challenge not only for general populations but also for athletes and physically active individuals who rely on optimal metabolic function for endurance, strength, and performance. As lifestyle factors play a pivotal role in diabetes prevention and management, investigating the influence of diet and stress on diabetes risk is crucial for optimizing long-term health outcomes, particularly in the context of sports medicine and rehabilitation (Fang et al., 2023). Diet and stress are two critical lifestyle factors influencing metabolic health, energy production, and recovery from physical exertion. Diet, particularly macronutrient composition, directly impacts glucose metabolism and insulin sensitivity. Excessive carbohydrate consumption can lead to postprandial hyperglycemia and insulin resistance, whereas protein and fat intake play essential roles in energy expenditure and muscle synthesis. Proper

glucose regulation is vital for endurance athletes, fitness enthusiasts, and individuals undergoing rehabilitation, making early detection of diabetes risk imperative for maintaining peak physical performance. Additionally, stress affects metabolic function in complex ways, with both positive and negative stress responses influencing glucose regulation. While positive stress (eustress), often experienced during training and competition, may enhance metabolic adaptation and insulin sensitivity, chronic negative stress (distress) can lead to excessive cortisol secretion, promoting insulin resistance and fat accumulation (Obermeyer & Emanuel, 2016; Organization, 2016). Understanding the bidirectional relationship between stress and metabolic health is essential for optimizing physical conditioning, recovery, and disease prevention. Advancements in machine learning provide a novel approach to identifying and analyzing diabetes risk factors by detecting complex interactions between diet, stress, and glucose metabolism. Unlike traditional statistical models, interpretable machine learning algorithms, such as Random Forest, allow for precise predictions while maintaining transparency in the decision-making process. These models offer valuable insights into how dietary intake and stress-related factors influence metabolic function, enabling healthcare professionals, sports scientists, and fitness experts to develop personalized interventions for diabetes prevention and physical performance optimization. By integrating dietary data and stress indicators into predictive models, this study aims to enhance the accuracy of diabetes detection while examining the relative impact of these factors on metabolic health (American Diabetes Association, 2014a; Chrousos, 2009). The primary objectives of this study are to assess the relationship between diet, stress, and diabetes risk using machine learning, examine the influence of carbohydrate consumption and stress variations on metabolic function, and evaluate how diabetes-related impairments affect exercise capacity and sports participation. Additionally, this study seeks to develop a predictive framework for early diabetes detection, offering practical applications for sports medicine, rehabilitation, and metabolic health management. The findings will be valuable for athletes, coaches, and healthcare professionals in designing personalized nutrition and training programs that enhance metabolic resilience and prevent diabetes-related physical impairments (Federation, 2019; Leo et al., 2023). This research is directly relevant to sports science and physical activity medicine, as it provides a deeper understanding of how diet and stress interact with metabolic health and exercise performance. By leveraging machine learning to identify diabetes risk factors (Hu, 2011; LeCun et al., 2015), this study contributes to the development of precision-based interventions that integrate nutrition, stress management, and structured exercise programs. Ultimately, bridging the gap between diabetes risk assessment and exercise-based interventions will support improved physical function, athletic performance, and long-term health outcomes for both athletes and individuals at risk of metabolic disorders. This study aims to fill these gaps by focusing on the interpretability of machine

learning models in diabetes detection. We integrate dietary and stress-related data to develop a comprehensive predictive model that addresses the shortcomings of current research. Additionally, we conduct a thorough analysis of feature importance and correlations to identify the most influential factors contributing to diabetes risk. By doing so, we aim to provide a more transparent and clinically applicable approach to diabetes detection, offering healthcare professionals actionable insights that can inform both preventive and therapeutic strategies. The main contributions of this paper are as follows:

- We present a novel machine learning model that integrates both dietary and stress-related features to enhance the prediction accuracy and interpretability of diabetes risk models.
- We conduct a comprehensive analysis of the importance of different dietary and stress features, revealing that dietary factors have a more significant impact on diabetes risk than stress factors, with positive stress showing a greater influence than negative stress.
- We demonstrate the practical application of this model in clinical settings, achieving high prediction accuracy while maintaining model interpretability, thereby providing a tool that can be effectively used by healthcare professionals.

The remainder of this paper is organized as follows: Section 2 provides a detailed overview of the related work in the fields of diabetes prediction and machine learning model interpretability. Section 3 describes the methodology, including data preprocessing, feature selection, and the machine learning models used. Section 4 presents the experimental results and analysis, highlighting the impact of dietary and stress factors on diabetes prediction. Section 5 discusses the implications of the findings and potential areas for future research. Finally, we are summarizing the key contributions and their relevance to clinical practice.

2. Related Word

2.1 Diabetes Detection

Diabetes mellitus, particularly type 2 diabetes, has become a major global public health issue, with its prevalence rapidly increasing in both developed and developing countries. Early detection of diabetes is crucial for effective management and prevention of complications. Traditionally, diabetes detection has relied on clinical tests such as fasting plasma glucose (FPG), oral glucose tolerance test (OGTT), and glycated hemoglobin (HbA1c) levels (American Diabetes Association, 2014b). However, these tests often require invasive procedures and may not be suitable for large-scale screening. In recent years, machine learning techniques have been increasingly applied to

diabetes detection, leveraging a wide range of clinical and non-clinical data, including demographic factors, lifestyle habits, and genetic markers (Choi et al., 2019; Kavakiotis et al., 2017). While these approaches have shown promise, most existing studies have primarily focused on using electronic health records (EHRs) and clinical biomarkers for prediction. For instance, logistic regression, support vector machines (SVM), and decision trees have been widely utilized, but these methods often lack interpretability, making it difficult for clinicians to understand the underlying factors driving the model's predictions (Faruque & Sarker, 2019). Moreover, while some studies have explored the use of deep learning models, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), these models, despite their high accuracy, are often criticized for their "black-box" nature, which poses significant challenges for interpretability in clinical settings (Shickel et al., 2017). These models tend to obscure the contribution of individual features, making it difficult to derive actionable insights about how diet, lifestyle, or other factors influence diabetes risk. A notable gap in the existing literature is the limited focus on dietary data as a primary predictive feature for diabetes. While some studies have incorporated lifestyle factors, very few have concentrated on diet as the central element of their models (Luo et al., 2016). Additionally, the combination of dietary and stress-related features in diabetes prediction has been sparsely explored, despite the well-documented impact of stress on metabolic health (Hackett & Steptoe, 2017). This represents a significant oversight, as integrating these factors could enhance the predictive power and practical relevance of machine learning models in real-world clinical applications. Another critical limitation of existing studies is the lack of comprehensive feature analysis, particularly in terms of understanding which specific dietary or lifestyle factors most strongly influence diabetes risk. Most machine learning models provide predictions without offering insights into the relative importance of each feature or the potential correlations between them (Lundberg, 2017). This limits the clinical applicability of these models, as healthcare providers require not only accurate predictions but also a clear underlying factors to make informed decisions. In contrast to these limitations, our study advances the field by focusing on the interpretability of machine learning models in diabetes detection. We integrate both dietary and stress-related data to develop a comprehensive prediction model, addressing the gaps in current research. Furthermore, we conduct a thorough feature importance analysis and correlation analysis to identify the most influential factors contributing to diabetes risk. By doing so, we aim to provide a more transparent and clinically applicable approach to diabetes detection, offering healthcare professionals actionable insights that can inform preventive and therapeutic strategies.

2.2 Diet and Lifestyle

Diet and lifestyle play critical roles in the development and management of type 2 diabetes. Numerous studies have established the link between dietary

patterns, lifestyle choices, and diabetes risk. Diets high in refined sugars, saturated fats, and processed foods are consistently associated with an increased risk of developing type 2 diabetes (Hu et al., 2001). Conversely, diets rich in whole grains, fiber, fruits, and vegetables are associated with a reduced risk (Montonen et al., 2003). In addition to diet, lifestyle factors such as physical activity, smoking cessation, and moderate alcohol consumption are also recognized as crucial in mitigating the risk of type 2 diabetes (Laaksonen et al., 2005; Willi et al., 2007). Despite the extensive research on diet and lifestyle factors, most studies have treated these factors in isolation, without fully considering the complex interplay between diet, lifestyle, and psychological factors, particularly stress. Stress, both chronic and acute, has been identified as a significant contributor to metabolic dysfunction and diabetes risk (Hackett & Steptoe, 2017). However, the distinction between different types of stress—positive (e.g., challenge-related stress) and negative (e.g., hindrance-related stress)—and their specific impacts on diabetes has not been thoroughly explored in the existing literature. The study by Gan et al. provides important insights into how different dimensions of job stress affect mental health (Gan, 2014), emphasizing the differential effects of positive and negative stress. Their meta-analysis revealed that while both types of stress have significant health implications, the mechanisms through which they affect health outcomes may differ. This distinction is crucial for diabetes research, as stress is a known modifiable risk factor for metabolic diseases, including type 2 diabetes (Lloyd et al., 2005). However, there has been a lack of research directly examining the effects of positive and negative stress on diabetes risk, particularly in the context of dietary and lifestyle factors. Our study seeks to address this gap by incorporating both dietary patterns and stress-related data, specifically differentiating between positive and negative stress, to develop a more comprehensive model for diabetes prediction. This approach not only enhances the predictive power of the model but also offers new insights into how these factors interact to influence diabetes risk. By doing so, we provide a more nuanced understanding of the roles that diet, lifestyle, and stress play in diabetes, highlighting the need for personalized intervention strategies that consider the full spectrum of these factors. In summary, while significant progress has been made in understanding the impact of diet and lifestyle on diabetes, there remains a critical need to explore how different types of stress interact with these factors to influence disease risk. Our study contributes to this emerging area of research by integrating dietary and stress-related data, offering a more holistic approach to diabetes prediction and management.

2.3 Interpretable Machine Learning

The study of Interpretable Machine Learning (IML) is of critical importance in the healthcare domain, particularly in disease diagnosis. In this context, the decision-making process of models must not only achieve high accuracy but also possess a high degree of interpretability. This interpretability

is directly tied to the trustworthiness and effectiveness of clinical decisions, as healthcare professionals and patients need to understand how a model arrives at its conclusions. This understanding ensures that these decisions are ethically sound and can be appropriately applied and validated in real-world scenarios (Molnar, 2020; Ribeiro et al., 2016). In some cases, interpretability may even outweigh predictive accuracy because it allows clinicians to identify and correct potential model biases, thereby avoiding misdiagnoses or inappropriate treatments (Macqueen, 1967). In this study, we explored three distinct machine learning approaches—K-means clustering, Random Forest classification, and Deep Neural Networks (DNN)—to assess their interpretability in the context of diabetes detection, particularly when incorporating diet and stress-related features. Each method not only represents a different learning paradigm but also reflects the key challenges that machine learning faces in healthcare: how to accurately capture underlying health risk factors from complex data while ensuring that the reasoning process of the model is understandable and trustworthy to medical practitioners. We compared these methods to evaluate their strengths and limitations, with the aim of providing more reliable machine learning solutions for disease diagnosis. K-means clustering is a classic unsupervised learning method. Although its direct interpretability is limited, it can still offer some level of explanation through the analysis of cluster centroids (Macqueen, 1967). The core of K-means lies in partitioning the data into several non-overlapping clusters, where analyzing the centroids can help researchers identify typical characteristics of each cluster (Jain, 2010). In the context of diabetes detection, these cluster centroids can be used to infer dietary and stress patterns associated with higher or lower risks (Lloyd, 1982). While K-means is inherently a "black-box" model, interpreting the cluster centroids can provide useful insights for clinical research (Murphy, 2012). In contrast, Random Forest is a supervised learning method with an inherent mechanism for feature importance evaluation, which grants it a high degree of interpretability (Breiman, 2001a). Random Forest operates by constructing a large number of decision trees, each contributing to the final outcome. Its feature importance metrics (e.g., Gini impurity or information gain) allow researchers to identify which features are most influential in the model's decision-making process (Breiman, 2001b). In this study, this characteristic of Random Forest enabled us to quantify the impact of dietary and lifestyle factors on diabetes risk, thereby supporting medical decision-making (Liaw, 2002). However, Deep Neural Networks (DNN), due to their complex architecture and highly nonlinear nature, are often regarded as "black-box" models that are difficult to interpret directly (Goodfellow, 2016). Despite typically excelling in predictive accuracy, the multiple layers of nonlinear transformations within DNNs make it challenging to understand their decision-making processes (Montavon et al., 2018). To address this challenge, several interpretability techniques, such as SHAP (SHapley Additive exPlanations) values and LIME (Local Interpretable Model-agnostic Explanations), have been proposed in

recent years (Lundberg, 2017; Ribeiro et al., 2016). These methods approximate the model's output to generate more interpretable local or global explanations, helping researchers understand which input features have the greatest influence on specific predictions (Shrikumar et al., 2017). In our study, although we employed DNN models combined with these interpretability techniques, the experimental results indicated that DNNs did not demonstrate a clear advantage in diabetes detection. We primarily used them as a comparison to evaluate their performance on the same dataset relative to other models like Random Forest and K-means (Sundararajan et al., 2017). In summary, K-means clustering provides limited interpretability through the analysis of cluster centroids, Random Forest offers direct interpretability through feature importance analysis, and while deep learning models are complex, they can still achieve a degree of interpretability through advanced explanation techniques. Each of these methods has its own strengths, collectively demonstrating how to enhance the transparency and interpretability of models while ensuring accuracy, thereby supporting clinical applications in diabetes detection (Tjoa & Guan, 2020).

3. Data and Interpretable Methodology

3.1 Data Source and Structure

The data for this study are sourced from the China Health and Nutrition Survey (CHNS), a longitudinal survey that provides comprehensive information on health, nutrition, and lifestyle across various provinces in China. The survey, a collaboration between the Carolina Population Center at the University of North Carolina and the Chinese Center for Disease Control and Prevention, is recognized for its extensive and diverse dataset, making it a valuable resource for public health research. This study utilizes dietary data from the CHNS Constructed Dietary Value dataset (c12diet) and stress-related data from the CHNS Physical Examination dataset (pstress_12). The dietary data include variables such as carbohydrate intake (D3CARBO), fat intake (D3FAT), calorie intake (D3KCAL), and protein intake (D3PROTN). Stress-related data include variables such as perceived stress (U551 to U564). For this analysis, we focused on dietary and stress variables in relation to diabetes diagnosis, as indicated by the U24A variable, shown in Table 1, with bolded text representing positive stress, and unbolded text representing negative stress.

Table 1: (a) The Key Features used in Our Analysis

FEATURE	DESCRIPTION
D3CARBO	3-Day Average: Carbohydrate Intake (g)
D3FAT	3-Day Average: Fat Intake (g)
D3KCAL	3-Day Average: Calorie Intake (kcal)
D3PROTN	3-Day Average: Protein Intake (g)
U551	Upset

Table 1: (b) The Key Features used in Our Analysis

FEATURE	DESCRIPTION
U552	Unable to Control Important Things
U553	Nervous and stressed
U554	Deal Successfully
U555	Effectively Cope With
U556	Confident About Your Ability
U557	Things Were Going Your Way
U558	Could Not Cope With
U559	Able To Control Irritations
U560	Were on Top of Things
U561	Angered
U562	Thinking About Things
U563	Able To Control the Way You Spend Your Time
U564	Difficulties
U24A	Diagnosed with Diabetes (Output)

3.2 Methodologies

In this section, we delve into the specific machine learning methodologies applied in this study, focusing on their implementation and interpretability aspects.

3.2.1 K-means Clustering: Interpretability through Cluster Analysis

K-means clustering is a widely used unsupervised learning algorithm that aims to reveal underlying structures in data by partitioning it into K clusters. In this study, we employed K-means clustering as the initial step to explore the impact of dietary and stress-related variables on diabetes detection. The interpretability of K-means primarily lies in analyzing the positions of cluster centers and the distribution of data points within each cluster. This information can help us understand which patterns of diet and stress are more closely associated with the risk of diabetes. The core idea behind the algorithm is to minimize the sum of the distances between data points and their corresponding cluster centers. For each data point x_i, we calculate its Euclidean distance to each cluster center $c_j$and assign the point to the nearest cluster. This process is represented by the following equation:

$$d(x_i, c_j) = \sqrt{\sum_{m=1}^{n}(x_{im} - c_{jm})^2} \quad (1)$$

where x_{im} represents the value of the m^{th} feature of data point x_i, and c_{jm} represents the value of the m^{th} feature of cluster center c_j. Through iterative updates, the algorithm adjusts each cluster center to minimize the total within-cluster variance, which is the sum of the squared distances from each

data point to its cluster center:

$$W(C) = \sum_{j=1}^{K} \sum_{x_i \in S_j} \|x_i - c_j\|^2 \quad (2)$$

The optimization process stops when the cluster centers converge, meaning that they no longer change significantly between iterations. The unique aspect of K-means lies in its intuitive clustering results; although often considered a "black box," we can infer the representative characteristics of each cluster by analyzing the positions of the cluster centers. Additionally, this study introduces the concept of positive and negative stress, allowing us to use K-means clustering to identify which types of stress have a greater impact on diabetes. Compared to traditional K-means analysis, this approach provides richer interpretability by integrating different types of stress into the clustering analysis.

3.2.2 Random Forest Classification: Interpretability through Decision Rules

Random Forest is a powerful supervised learning algorithm that constructs multiple decision trees to perform classification or regression tasks. Its interpretability is derived not only from feature importance analysis but also from the examination of individual decision tree rules. The widespread application of Random Forest in the medical field is partly due to its ability to provide a transparent decision-making process, which is crucial for clinical applications. Each decision tree in the Random Forest is built on a bootstrapped sample of the training data. For each node, the algorithm selects a feature to split the data based on criteria such as maximizing information gain or minimizing Gini impurity. Gini impurity $Gini(t)$ is a commonly used metric to assess the purity of a node and is calculated as follows:

$$Gini(t) = 1 - \sum_{i=1}^{C} p_i^2 \quad (3)$$

where p_i represents the proportion of samples belonging to class *i* at node *t*. By minimizing Gini impurity, the decision tree selects the feature that best separates the different classes, leading to a clear decision path. Another commonly used splitting criterion is information gain, which measures how much a feature reduces the uncertainty at a node. The information gain is calculated as follows:

$$IG(t, f) = H(t) - \sum_{v \in V(f)} \frac{|t_v|}{|t|} H(t_v) \quad (4)$$

where $H(t)$ is the entropy at node *t*, $V(f)$ represents the possible values of feature *f*, and t_v denotes the set of child nodes after splitting by *f*. Higher information gain indicates that the feature is more effective at classifying

the data, making it a better candidate for splitting. Random Forest combines the predictions of multiple decision trees to improve model stability and accuracy. The final output is determined by majority voting (for classification) or averaging (for regression) the predictions of all trees. This ensemble approach reduces the risk of overfitting associated with individual decision trees and provides more robust predictions. In Random Forest, feature importance is quantified by evaluating the contribution of each feature to the model's decisions across all trees. This can be achieved by summing the reductions in Gini impurity caused by each feature:

$$FI(f) = \frac{1}{T}\sum_{t=1}^{T} \Delta\, Gini_t(f) \quad (5)$$

where $FI(f)$ represents the importance of feature f, T is the total number of trees, and $\Delta Gini_t(f)$ is the reduction in Gini impurity caused by feature f in tree t. In this study, we enhanced interpretability by not only performing traditional feature importance analysis but also by examining the decision paths within individual trees in the Random Forest. This dual approach provided deeper insights into how specific features, such as dietary and stress variables, contribute to diabetes prediction. Such transparency is invaluable in clinical decision-making, offering healthcare professionals concrete guidance based on the model's internal reasoning.

4. Experimental Results and Analysis

In this section, we conducted two main experiments to evaluate the predictive power of our models. The first experiment utilized only dietary structure data, while the second experiment incorporated both positive and negative stress data alongside the dietary data. The results indicate that adding stress data significantly improves the model's accuracy and generalizability, particularly in capturing the nuanced effects of positive stress on diabetes risk. Given the imbalance in our dataset—where the number of diabetes patients was significantly lower than that of healthy individuals—we applied data preprocessing techniques, including up sampling and down sampling, to balance the dataset. Additionally, all features were normalized to ensure consistency across the models. For transparency and reproducibility, all code used in these experiments is available upon request from the corresponding authors.

4.1 Analysis of Results Only Based on Dietary Structure

In this section, we analyze the performance of the model when only dietary structure features are used for diabetes detection. The features considered in this analysis include protein intake (d3protn), carbohydrate intake (d3carbo), fat intake (d3fat) and calorie intake (d3kcal). Our results show that the model achieved an impressive prediction accuracy of 99.67% as shown in

Table 2 and 3, indicating that dietary structure alone can effectively distinguish between diabetic and non-diabetic individuals. This high accuracy suggests that the dietary patterns captured by these features are strongly associated with the risk of diabetes.

Table 2: Classification Report for Diabetes Prediction Using Random Forest

CLASS	PRECISION	RECALL	F1-SCORE	SUPPORT
NO DIABETES (0)	1.00	0.99	1.00	18,708
DIABETES (1)	0.99	1.00	1.00	17,816
ACCURACY			99.67%	36,524

Table 3: Confusion Matrix of Diabetes Prediction Using Random Forest

CLASS	NO DIABETES (0)	DIABETES (1)
NO DIABETES (0)	18708	120
DIABETES (1)	0	17816

Feature Importance: Among the four dietary features analyzed, carbohydrate intake (d3carbo) was identified as the most influential factor as shown in Fig.1, with the highest importance score of 0.2791. This result highlights that carbohydrate consumption plays a more significant role in diabetes detection compared to the other three dietary factors. The other features, including protein, fat, and calorie intake, also contributed to the model's predictions but to a lesser extent.

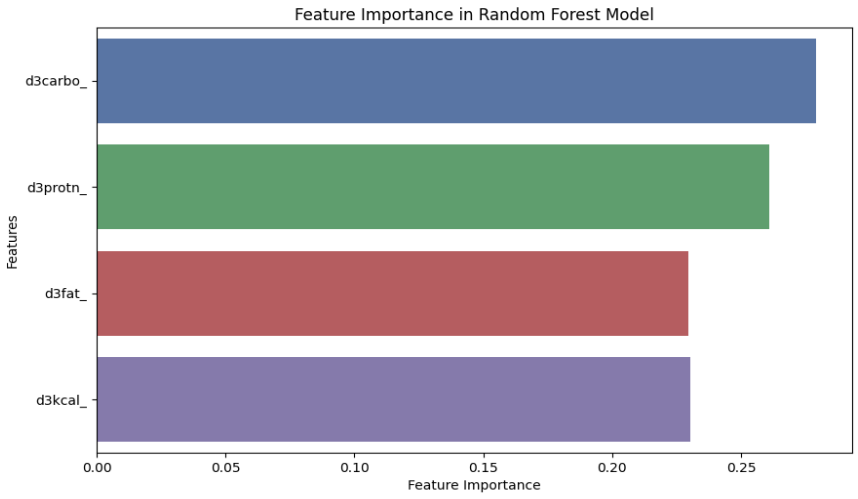


Figure 1: Feature Importance in Random Forest Model Based on Dietary Structure

Correlation Analysis: To explore the relationships between the four dietary features, we conducted a correlation analysis. The results, as shown in Fig. 2, indicate that while some features exhibit moderate correlations, none of these correlations are statistically significant (all p-values > 0.05). This lack of significant correlation suggests that each dietary feature contributes independently to the model, providing unique and non-redundant information.

This independence among features enhances the robustness of the feature importance rankings, ensuring that the model's predictions are not driven by redundant or highly correlated inputs.

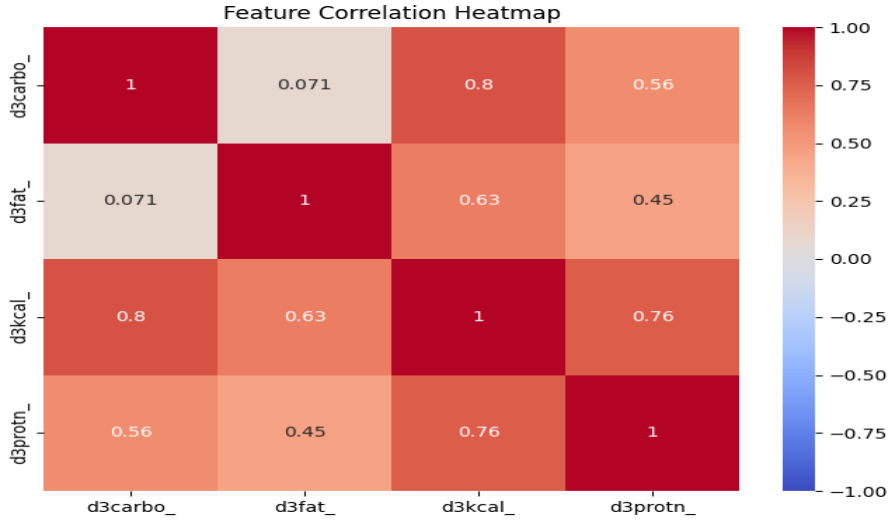
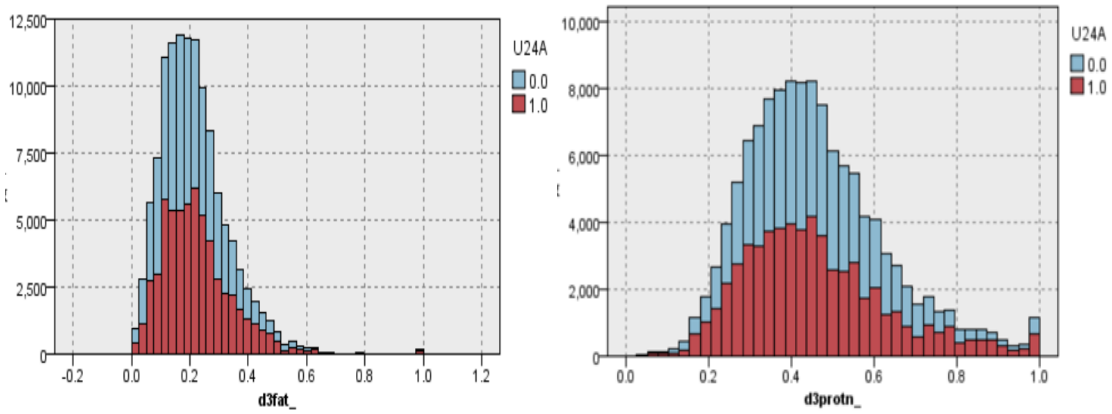
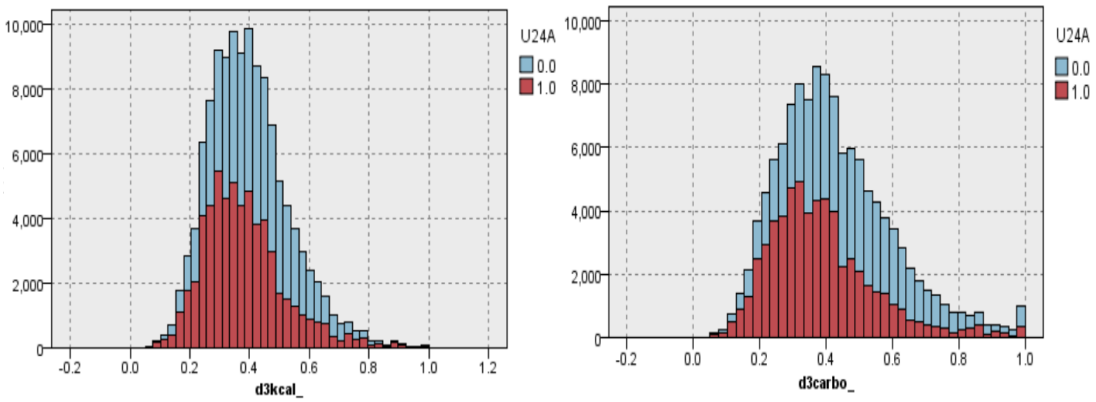


Figure 2: Feature Correlation Heatmap in Random Forest Model Based on Dietary Structure

Model Interpretability: Although the model based solely on dietary structure achieved high predictive accuracy, the decision-making process behind these predictions remains complex. As shown in Fig. 3, the features overlap between the two outcome classes, which further complicates the understanding of the model's decision pathways. Even though the random forest model relies on just four features, the rules it generates are still difficult to interpret directly, as illustrated in Fig. 4. This complexity highlights a common trade-off in machine learning models: as accuracy improves, interpretability often diminishes. In this case, while we can trust the model's predictions, understanding the exact decision-making process the model follows remains challenging.



(a) d3fat (b) d3protn



(c) d3kcal (d) d3carbo

Figure 3: Sample Distribution of Four Features

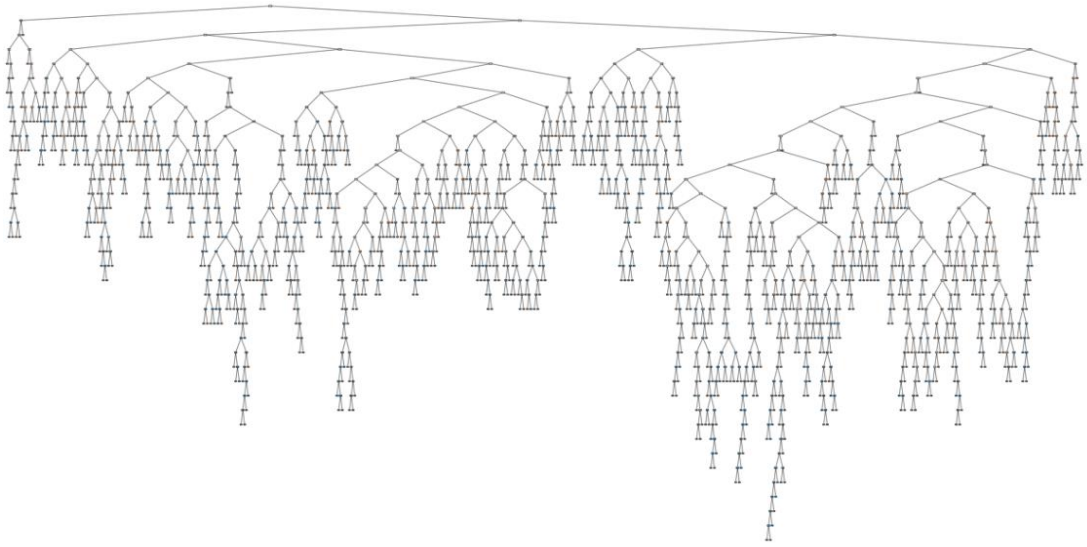


Figure 4: Multiple Decision Tree Rules for Random Forest

In contrast, the DNN model trained on the same dietary features performed poorly in comparison, further underscoring the strength of the Random Forest model in capturing the relevant dietary patterns. However, the inability of the deep learning model to outperform Random Forest in this scenario suggests that more complex models do not necessarily lead to better outcomes when interpretability and limited feature sets are critical.

Table 4: Accuracy On DNN Model in Different Datasets

DATASETS	DIETARY STRUCTURE	DIETARY STRUCTURE+ LIFE STRESS
ACCURACY	70.85%	95.04%

The analysis demonstrates that carbohydrate intake is a key factor in predicting diabetes, and dietary structure alone provides a strong basis for diabetes detection. In the next section, we will explore whether combining dietary features with stress-related data can enhance both the predictive power and interpretability of the model.

4.2 Analysis of Results Added with Life Stress

In this section, we introduce the inclusion of life stress features into our predictive models. The stress features, as shown in Table 1, are categorized into positive and negative stress. By incorporating these stress features, we not only achieve improved results in the Random Forest model but also enable easier and more accurate predictions of diabetes in neural networks in Fig.4 and K-means clustering models in Fig.5. The addition of stress features enhances the interpretability of the models, particularly in capturing complex relationships that were previously difficult to discern. This improvement allows for the effective use of non-invasive, simple daily features to predict at-risk populations early, facilitating timely intervention.

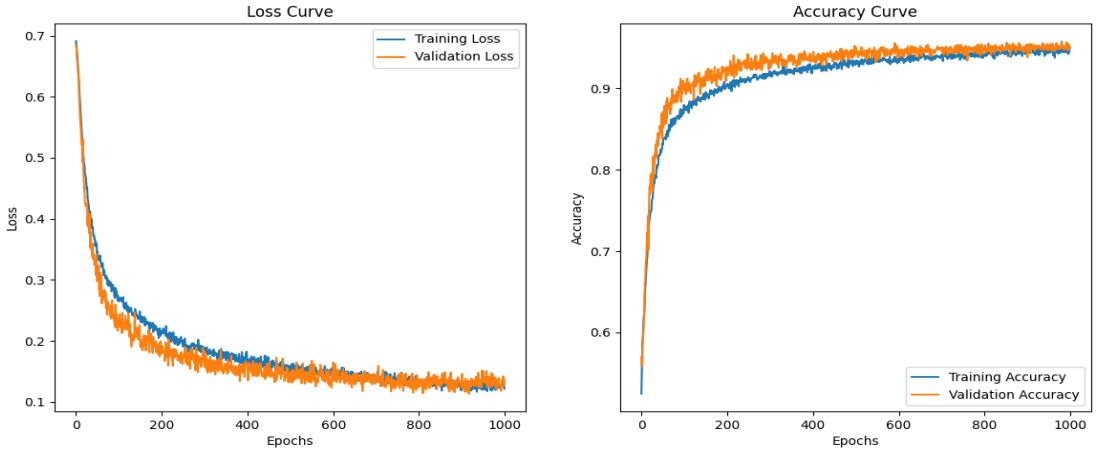


Figure 4: Performance of DNN Mode in the Added with Life Stress

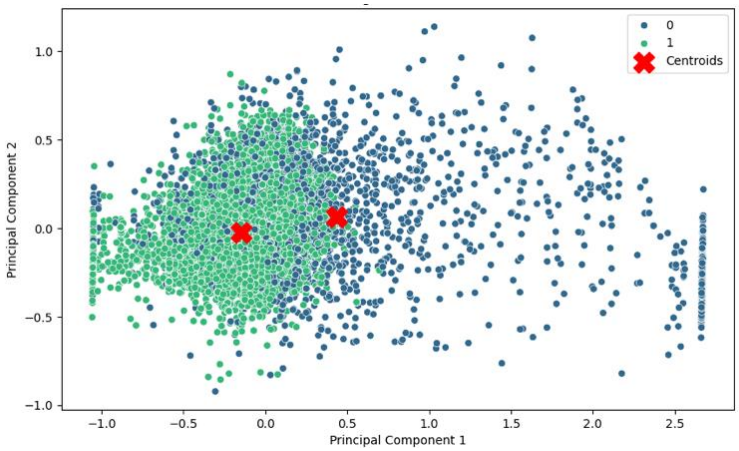


Figure 5: Classification Results after PCA Reduction in K-means Clustering

Moreover, we conducted a correlation analysis to explore the relationship between stress and dietary features. The results indicate a significant correlation between stress and dietary habits. Specifically, dietary features are negatively correlated with stress, suggesting that certain dietary patterns may help mitigate stress levels, or conversely, that higher stress levels may be associated with reduced intake of specific dietary components.

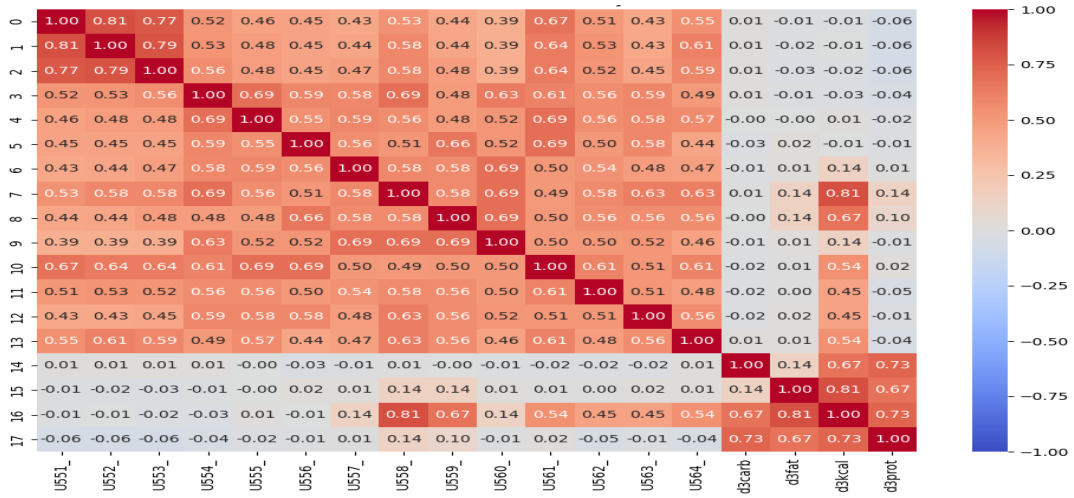


Figure 6: Correlation Matrix of Dietary and Stress Features

This negative correlation implies a potential bi-directional influence, where managing diet could serve as a strategy to alleviate stress, and vice versa. These findings highlight the importance of considering both dietary and stress-related factors in predictive models for diabetes, as they may provide valuable insights for early detection and prevention strategies.

4.3 Analysis of Importance of Dietary and Stress Features

In this section, we examine the importance of various dietary and stress features in predicting diabetes within our models. As illustrated in the accompanying feature importance chart in Fig. 7, dietary features (represented by blue bars) exhibit greater importance than stress features in our models. Notably, among the stress features, positive stress (represented by green bars) is more influential than negative stress (represented by red bars). These finding challenges traditional assumptions, where negative stress is typically seen as the primary contributor to adverse health outcomes, including diabetes.

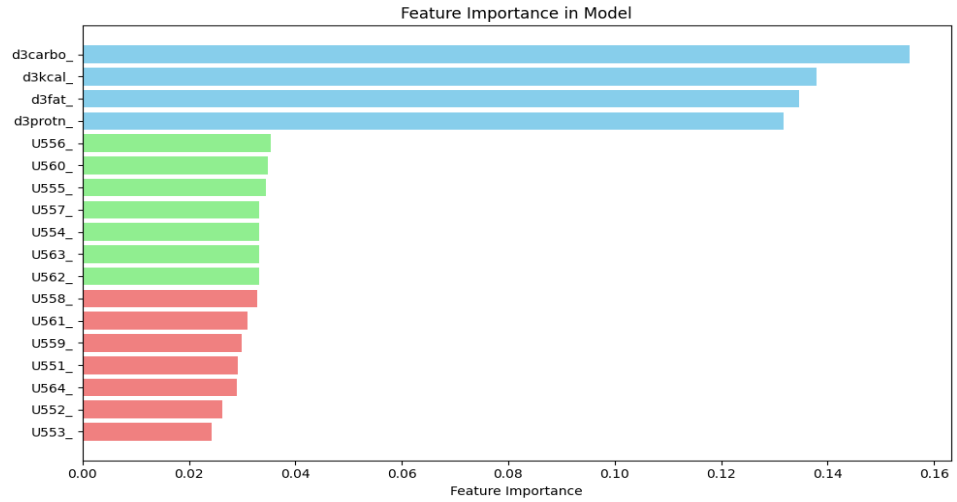


Figure 7: The Effect of Diet and Stress on Diabetes Detection

The importance of these features was assessed using a neural network model, where input variables are processed through multiple layers of neurons, each connected by weights. The magnitude of these weights determines the contribution of each feature to the final prediction—larger weights indicate a greater impact. Our analysis reveals that dietary features, such as carbohydrate intake (d3carbo), caloric intake (d3kcal), fat intake (d3fat), and protein intake (d3protn), significantly influence the model's predictions. These findings highlight the critical role of dietary patterns in diabetes risk, likely due to their direct effects on metabolism and blood glucose levels. Interestingly, positive stress indicators (green bars) exert a stronger influence on diabetes prediction than negative stress indicators (red bars). This result is counterintuitive to the traditional view that primarily associates negative stress with poor health outcomes. The significant impact of positive stress might suggest that stress related to motivation or challenges could influence physiological processes associated with diabetes in ways not previously understood, potentially affecting lifestyle choices that contribute to diabetes risk. In conclusion, the analysis underscores the dominant role of dietary habits in predicting diabetes risk, while stress factors, particularly positive stress, also play a significant role. The unexpected finding that positive stress is more influential than negative stress suggests a need to revisit traditional perspectives on stress and health, opening new avenues for research and potential interventions in diabetes prevention and management.

5. Conclusion

This study underscores the significant role of dietary habits and stress in diabetes risk and their implications for physical activity, sports performance, and metabolic health. By leveraging interpretable machine learning models, we identified the complex interactions between carbohydrate intake, stress responses, and glucose metabolism, offering new insights into early diabetes detection and prevention strategies tailored to active individuals and athletes. Our findings highlight that diet, particularly carbohydrate consumption, is the most critical determinant of diabetes risk, reinforcing the importance of nutritional regulation in optimizing metabolic function, endurance, and physical performance. Additionally, the impact of stress on diabetes risk challenges conventional assumptions, with positive stress exhibiting a stronger influence on metabolic dysfunction than negative stress, suggesting that adaptive stress responses play a crucial role in physiological adaptation and energy regulation. The practical implications of these findings are highly relevant to sports science, rehabilitation, and exercise physiology. Diabetes-related impairments in glucose metabolism, insulin sensitivity, and muscle function can significantly hinder an individual's ability to engage in regular physical activity, leading to reduced exercise tolerance, slower recovery, and diminished performance in sports and daily activities. Understanding the interactions between diet, stress, and metabolic health allows for the development of personalized interventions

that combine structured exercise regimens, dietary modifications, and stress management strategies to prevent and manage diabetes more effectively. Athletes, fitness professionals, and sports medicine practitioners can apply these insights to enhance training programs, improve recovery protocols, and support long-term metabolic health in individuals at risk of diabetes. Future research should focus on longitudinal studies assessing the impact of dietary adjustments and stress regulation on metabolic resilience in physically active populations. Additionally, investigating how different types of exercise, including endurance training, resistance training, and high-intensity interval training (HIIT), influence glucose regulation and diabetes risk could provide further evidence for integrating exercise-based interventions into metabolic health management. By bridging the gap between machine learning, nutritional science, stress physiology, and sports medicine, this study contributes to the development of precision-based strategies for diabetes prevention and performance optimization. As the integration of technology, data-driven diagnostics, and personalized medicine continues to advance, the potential for improving metabolic health, enhancing physical endurance, and preventing diabetes-related complications through targeted lifestyle interventions becomes increasingly promising. Healthcare professionals, sports scientists, and fitness experts must continue to explore innovative, data-driven approaches to ensure that individuals at risk of diabetes can maintain optimal physical function, participate in sports, and lead healthier, more active lives.

Acknowledgments

This research was supported by the Beijing Genomics Institute under grant numbers 62176164 and 62203134. The author gratefully acknowledges the support provided by the "Design of phage tail protein based on adversarial machine learning" project and the "Construction of artificial phage model based on deep generative adversarial network learning" project. And the datasets is provided by China Health and Nutrition Survey (CHNS) in https://www.cpc.unc.edu/projects/china/data/datasets.

References

American Diabetes Association. (2014a). Diagnosis and classification of diabetes mellitus. *Diabetes care*, *37*(Supplement_1), S81-S90.

American Diabetes Association. (2014b). Standards of medical care in diabetes—2014. *Diabetes care*, *37*(Supplement_1), S14-S80.

Breiman, L. (2001a). Random forests. *Machine learning*, *45*, 5-32.

Breiman, L. (2001b). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, *16*(3), 199-231.

Choi, B. G., Rha, S.-W., Kim, S. W., Kang, J. H., Park, J. Y., & Noh, Y.-K. (2019). Machine learning for the prediction of new-onset diabetes mellitus during 5-year follow-up in non-diabetic patients with cardiovascular risks.

Yonsei medical journal, *60*(2), 191.

Chrousos, G. P. (2009). Stress and disorders of the stress system. *Nature reviews endocrinology*, *5*(7), 374-381.

Fang, S., Guo, C., Chen, R., Chen, Y., & Xu, C. (2023). Observing the Clinical Outcomes of Single Port Endoscopic Posterolateral TLIF in Retired Athletes. *Revista multidisciplinar de las Ciencias del Deporte*, *23*(91).

Faruque, M. F., & Sarker, I. H. (2019). Performance analysis of machine learning techniques to predict diabetes mellitus. 2019 international conference on electrical, computer and communication engineering (ECCE),

Federation, I. D. (2019). Idf diabetes atlas. 2013. *International Diabetes Federation*.

Gan, Y. e. a. (2014). The relationship between job stress and mental health: A meta-analysis of studies assessing job stress by the Job Content Questionnaire and its implications for job stress management. *World Journal of Medical Sciences*, *4*(3), 100-106.

Goodfellow, I. (2016). Deep learning. In: MIT press.

Hackett, R. A., & Steptoe, A. (2017). Type 2 diabetes mellitus and psychological stress—a modifiable risk factor. *Nature reviews endocrinology*, *13*(9), 547-560.

Hu, F. B. (2011). Globalization of diabetes: the role of diet, lifestyle, and genes. *Diabetes care*, *34*(6), 1249-1257.

Hu, F. B., Manson, J. E., Stampfer, M. J., Colditz, G., Liu, S., Solomon, C. G., & Willett, W. C. (2001). Diet, lifestyle, and the risk of type 2 diabetes mellitus in women. *New England Journal of Medicine*, *345*(11), 790-797.

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern recognition letters*, *31*(8), 651-666.

Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research. *Computational and structural biotechnology journal*, *15*, 104-116.

Laaksonen, D. E., Lindstrom, J., Lakka, T. A., Eriksson, J. G., Niskanen, L., Wikstrom, K., Aunola, S., Keinänen-Kiukaanniemi, S., Laakso, M., & Valle, T. T. (2005). Physical activity in the prevention of type 2 diabetes: the Finnish diabetes prevention study. *Diabetes*, *54*(1), 158-165.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436-444.

Leo, F., López-Gajardo, M., García-Calvo, T., Pulido, J., & González-Ponce, I. (2023). Adaptation to spanish and validation of the sport team socialization tactics questionnaire. *Revista multidisciplinar de las Ciencias del Deporte*, *23*(91).

Liaw, A. (2002). Classification and regression by randomForest. *R news*.

Lloyd, C., Smith, J., & Weinger, K. (2005). Stress and diabetes: a review of the links. *Diabetes spectrum*, *18*(2), 121-127.

Lloyd, S. (1982). Least squares quantization in PCM. *IEEE transactions on information theory*, *28*(2), 129-137.

Lundberg, S. (2017). A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874*.

Luo, W., Phung, D., Tran, T., Gupta, S., Rana, S., Karmakar, C., Shilton, A., Yearwood, J., Dimitrova, N., & Ho, T. B. (2016). Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. *Journal of medical Internet research*, *18*(12), e323.

Macqueen, J. (1967). Some methods for classification and analysis of multivariate observations. Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability/University of California Press,

Molnar, C. (2020). *Interpretable machine learning*. Lulu. com.

Montavon, G., Samek, W., & Müller, K.-R. (2018). Methods for interpreting and understanding deep neural networks. *Digital signal processing*, *73*, 1-15.

Montonen, J., Knekt, P., Järvinen, R., Aromaa, A., & Reunanen, A. (2003). Whole-grain and fiber intake and the incidence of type 2 diabetes. *The American journal of clinical nutrition*, *77*(3), 622-629.

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.

Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the future—big data, machine learning, and clinical medicine. *New England Journal of Medicine*, *375*(13), 1216-1219.

Organization, W. H. (2016). Global report on diabetes.

Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. *New England Journal of Medicine*, *380*(14), 1347-1358.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). " Why should i trust you?" Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining,

Shickel, B., Tighe, P. J., Bihorac, A., & Rashidi, P. (2017). Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. *IEEE journal of biomedical and health informatics*, *22*(5), 1589-1604.

Shrikumar, A., Greenside, P., & Kundaje, A. (2017). Learning important features through propagating activation differences. International conference on machine learning,

Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. International conference on machine learning,

Tjoa, E., & Guan, C. (2020). A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE transactions on neural networks and learning systems*, *32*(11), 4793-4813.

Willi, C., Bodenmann, P., Ghali, W. A., Faris, P. D., & Cornuz, J. (2007). Active smoking and the risk of type 2 diabetes: a systematic review and meta-analysis. *Jama*, *298*(22), 2654-2664.