Austrian Journal of Statistics
2024, Volume 53, 90–111.
http://www.ajs.or.at/
doi:10.17713/ajs.v53i5.1939

Mediation Analysis of Diabetes and Heart Diseases Influenced by Obesity Using Machine Learning Classifiers

Ajay Verma
Mathematics Division,
School of Advanced Sciences
and Languages,
VIT Bhopal University

Manisha Jain
Mathematics Division,
School of Advanced Sciences and
Languages,
VIT Bhopal University

Abstract

Purpose: This study examines the impact of body mass index on diabetes and heart disease among Indians. Multi-morbidity ailments are associated with diabetes. To understand the relationship between diabetes, body mass index, and heart disease, a study is undertaken.

Methods: The present study established a relationship between diabetes, heart disease and body mass index using mediation analysis and machine learning classifiers. R software with Hayes Macro Process and Python was used as a statistical tool to conclude the study.

Results: As a result of the present findings, the body mass index mediates the relationship between diabetes and heart disease and cannot be countered. This study's indirect impact is 4.6231, statistically significant at a 95% significance of the indirect effect of diabetes on heart disease is evident, as (BootLLCI), and (BootULCI) are both positive and do not contain zero. This indicates that there is a substantial mediation effect present. In classification, the TensorFlow classifier shows 99% accuracy and 97% precision, while the Linear S.V.C., NuSVC and Logistic Regression have an accuracy of 98%, 96% and 97%, which shows that the machine learning classifiers are more significant for the study

Conclusion: Our study examines how Body Mass Index (BMI) mediates diabetes and heart disease, which are statistically significant. Despite the close relationship between heart disease and diabetes, little is known about the pathways involved. Machine Learning Classifiers show that the risk of diabetes, heart disease and other diseases increases due to decreased body mass index.

Keywords: mediation, lifestyle factors, cardiovascular disease, obesity, comorbidities.

1. Introduction

In the realm of health issues, non-communicable diseases (N.C.D.s) persist as a risk to the overall welfare of society. Among these, diabetes and heart disease have emerged as pressing concerns. The escalating prevalence of these conditions necessitates a nuanced understanding of the underlying factors contributing to their onset and progression. BMI, a widely used metric for assessing body weight concerning height, has been identified as a potentially crucial play in the complex interplay of factors influencing diabetes and heart disease. India, undergoing rapid socioeconomic transitions, is witnessing substantial lifestyle and dietary patterns shifts, contributing to a growing health burden. This research explores the intricate relationship between BMI and the prevalence of diabetes and heart disease in India. However, recognizing the association between BMI and these health outcomes is likely multifaceted, and we propose a sophisticated analytical approach—regression-based mediation analysis—to unravel the underlying mechanisms and pathways through which BMI influences the incidence and progression of diabetes and heart disease.

Blood sugar levels in both men and women (>140 mg/dl) are around 14%, according to National Family Health Survey (NFHS-5, 2019-2021) (Maiti, Akhtar, Upadhyay, and Mohanty 2023). The World Health Organization (WHO) predicts that during the next 20 years, the population of people with type II diabetes mellitus will rise dramatically in developed and developing nations. The rise is predicted to be 46% in developed countries (55 million in 2000, 83 million in 2030) and 150% in developing countries (30 million in 2000, 80 million in 2030). A serious public health and medical problem, diabetes mellitus is thought to cause 4.6 million deaths every year around the world. India is going through a slow-motion diabetes emergency. The community must be aware of and educated about diabetes; social scientists working in community health are essential to this effort (Rathmann and Giani 2004). Epidemically, diabetic complications are classified as critical non-communicable diseases (N.C.D.s), such as heart disease (Murugesan, Snehalatha, Shobhana, Roglic, and Ramachandran 2007).

From a statistical point of view, its prevalence has impacted various sociodemographic groups (Shetty, Jena, and Kadithi 2013). Glucose levels in diabetes mellitus are abnormally high. More than 62 million people have been diagnosed with diabetes worldwide, making India the nation with the highest number of diabetics and heart disease Kaveeshwar and Cornwall (2014). It is estimated that in India, the number of people with diabetes and heart disease mellitus will triple by 2030 (Joshi 2015). Multifactorial factors, including genetics, the environment, and alterations in lifestyle, all have a role in the development of diabetes. There have been many articles on the prevalence of diabetes in India. So far, the most minor research has been done on this subject. India's rural people experience a greater rate of diabetes than its urban ones (Deepa, Bhansali, Anjana, Pradeepa, Joshi, Joshi, Dhandhania, Rao, Subashini, Unnikrishnan *et al.* 2014). One theory is that south Indians are the host people and north Indians are the migrating Asian populations. Urban locations are more likely to have access to trustworthy screening techniques and anti-diabetic drugs.

In contrast, rural areas are underserved regarding health care and access to preventative treatments. Significant advantages, such as improved treatment adherence and reduced complications, have also been linked to greater awareness about diabetes and its consequences (Rani, Raman, Subramani, Perumal, Kumaramanickavel, and Sharma 2008). Diabetes is multifactorial in India and is influenced by environmental and genetic variables, such as obesity brought on by rising living standards, constant urban migration, and lifestyle changes (Wells, Pomeroy, Walimbe, Popkin, and Yajnik 2016). Despite the high frequency of diabetes within the nation, the prevalence of the disease and its complications is generally modest in India. However, little national and multi-centric research has been done on the topic. The extrapolation of regional data may lead to incorrect projections for the entire country due to the variability of the Indian people in terms of culture, ethnicity, and socioeconomic conditions. There hasn't been much research in India on obesity as a risk factor for diabetes despite it being one of the significant risk factors. While global averages for BMI and diabetes have climbed over the past three decades, blood pressure and cholesterol have dropped or stayed the same Lakshminarayan and Tejaswi (2014). High BMI is an important cardiovascular disease risk factor and raised blood pressure, cholesterol, and glucose partly mediate its effects (Poirier, Giles, Bray, Hong, Stern, Pi-Sunyer, and Eckel 2006).

2. Related works

Seidell, Hautvast, and Deurenberg (1989) conducted a study that found that a waist-hip ratio (W.H.R.) of more than 1 for men and more significant than 0.85 for women is considered truncal obesity. Collaboration *et al.* (2009) have investigated the relationship between BMI and the risk of developing diabetes and heart disease. They conducted a collaborative analysis of 57 prospective studies and found that a higher BMI is associated with an increased risk of vascular diseases, including coronary heart disease and stroke. (Dudina, Cooney, Bacquer, Backer, Ducimetière, Jousilahti, Keil, Menotti, Njølstad, Oganov *et al.* 2011) Conducted a collaborative analysis of 58 prospective studies and found that BMI and abdominal adiposity are independently associated with an increased risk of cardiovascular disease. These findings suggest that considering overall BMI and abdominal adiposity is essential in assessing the risk of cardiovascular disease. Moreover, the molecular and metabolic mechanisms underlying cardiac dysfunction in diabetes have been explored. Despite not directly indicating adiposity, BMI is frequently used to assess obesity Shah and Braverman (2012). When the BMI exceeded 30 kg/m^2, it was considered obese; when it exceeded 25 kg/m^2, it was considered overweight (Gierach, Gierach, Ewertowska, Arndt, and Junik 2014). Males with waist circumferences (W.C.) $>$ 94 cm and females with W.C.$>$ 80 cm were categorized as having central and abdominal obesity.

Wang, Wang, Yang, Zhao, and Kuang (2015) conducted a pooled analysis of 97 prospective cohorts and found a positive association between BMI and coronary heart disease and stroke. These findings suggest that BMI plays a significant role in developing these diseases (Hruby and Hu 2015). Medically significant differences between lean and obese are somewhat arbitrary because body weights change continually between groups. (Farooqi, Khunti, Abner, Gillies, Morriss, and Seidu 2019) investigated that BMI and other factors, such as comorbid depression, have been found to increase the risk of cardiac events and mortality in people with diabetes. Conducted a systematic review and meta-analysis and found that the comorbid occurrence of diabetes and depression is associated with an increased risk of cardiovascular endpoints, including cardiovascular mortality, coronary heart disease, and stroke.

Furthermore, the role of mitochondrial dysfunction in diabetes-related cardiac pathogenesis has been studied, which has highlighted the involvement of mitochondrial dysfunction in the development of diabetic heart disease.

Adipose tissue that has developed in excess is considered obese (Longo, Zatterale, Naderi, Parrillo, Formisano, Raciti, Beguinot, and Miele 2019). Numerous epidemiological studies have demonstrated that when BMIs are below 25, morbidity rates slowly increase for all causes, including metabolic, cancer, and cardiovascular reasons (Tantengco 2022). As a result, a BMI of 25 to 30 should be regarded as medically essential and warrant therapeutic attention, mainly when linked to other risk factors like hypertension and glucose intolerance. de Lucia, Metzinger, and Wallner (2023) Reviewed the emerging pathways involved in diabetic cardiomyopathies and highlighted the importance of understanding the molecular and metabolic mechanisms contributing to cardiac dysfunction in diabetes. Obesity serves as the mediator between diabetes and heart disease. Machine learning classifiers have enhanced Traditional mediation analysis methods, allowing for a more accurate and nuanced understanding of these relationships. Integrating TensorFlow, NuSVC, and logistic regression in mediation analysis represents a novel approach to health research, offering the potential to uncover hidden patterns and interactions. S.V.C.s are particularly effective in high-dimensional spaces and are robust to overfitting, especially in cases where the number of dimensions exceeds the number of samples. Joachims and Joachims (2002) introduced the support vector machine (SVM), laying the foundation for S.V.C.s. Studies like that of Lutz, Zwygart, Thomann, Stucki, and Burla (2022) have used S.V.C.s to classify health-related data, showcasing their utility in distinguishing between different disease states based on clinical parameters. U-support vector Classification (NuSVC) is a variation of the traditional S.V.C., allowing more flexibility in specifying the number of support vectors and margin errors. Ketabchi, Moosaei, Razzaghi, and Pardalos (2019) introduced NuSVC, providing an alternative to the traditional C-SVC with better control over the model's complexity. NuSVC has been applied in various health studies to classify diseases and predict risks. For instance, Gupta, Kumar, Arora, and Raman (2022) used NuSVC to identify biomarkers for cardiovascular diseases. Abbott (1985) introduced logistic regression, which has been extensively used in medical research. Kleinbaum, Dietz, Gail, Klein, and Klein (2002) provide a comprehensive overview of logistic regression applications in health sciences. In obesity, diabetes, and heart disease, logistic regression helps identify key risk factors and their interactions. Wilson, Lorenz, Wilson, and Lorenz (2015) used logistic regression to analyze Framingham Heart Study data, identifying obesity as a significant predictor of diabetes and cardiovascular diseases. Recent advancements have demonstrated the potential of this integration. Nguyen, Ogburn, Schmid, Sarker, Greifer, Koning, and Stuart (2023) combined mediation analysis with machine learning techniques to explore the indirect effects of physical activity on cardiovascular health through body mass index (BMI). Their study highlighted the enhanced explanatory power and accuracy achieved through this combined approach.

3. Research objective

The primary aim is to discern the pathways through which BMI may impact the occurrence of diabetes and heart disease, shed light on potential covariates mentioned in the analysis [Figure 1] and contribute insights that can inform preventive strategies and healthcare interventions in the Indian population by using mediation analysis and machine learning classifiers.

4. Materials and methods

The straightforward relationship between X and Y is commonly known as the total effect of X on Y, as illustrated in Figure 1. We use the notation 'c' to represent this total effect, distinguishing it from c to c', which denotes the direct impact of X on Y after accounting for the influence of M. The formal analysis heuristic frequently employed to identify simple mediation effects follows directly from Baron and Kenny's (1986) definition of a mediator.

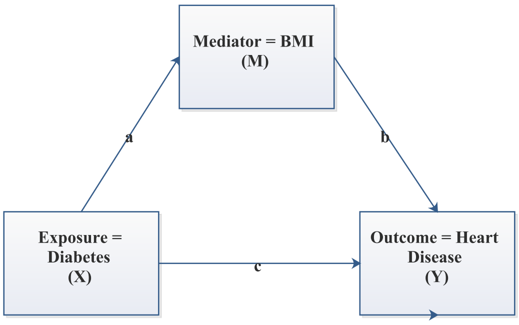

Figure 1: Road map of mediation model

According to this definition, variable M qualifies as a mediator if three conditions are met: X significantly predicts Y (i.e., $c \neq 0$), and X predicts M (i.e., $a \neq 0$ in Figure 1) M significantly predicts Y while controlling for X (i.e., $b \neq 0$). Baron and Kenny outline various analyses that should be conducted, and the results are evaluated based on these criteria. The assessment involves estimating the following equations:

$$Y = i_1 + cx + e_1$$

$$Y = i_2 + c'X + b_1M_1 + b_2M_2 + e$$

$$M = i_3 + a_1X + e_3$$

An intercept coefficient, denoted as ‘i’, signifies the point where the impact of variable X on Y diminishes to zero when variable M is introduced. This state, identified by James and Brett in 1984 as complete mediation, occurs when $X's$ effect on Y is entirely mediated by M. In cases where the effect of X on Y diminishes significantly but does not reach zero, it is termed partial mediation.

Two additional assumptions must be satisfied based on Baron and Kenny’s criteria for asserting mediation beyond meeting the conditions above. These include the absence of measurement error in M and the absence of a causal relationship where Y influences M." The basic mediation model using within-group centring is:

$$M_{ij} = dM_j + a_j \left(X_{ij} - \bar{X}_{.j} \right) + e_{ij}$$

$$Y_{ij} = d_{Yj} + c'_j \left(X_{ij} - \bar{X}_{.j} \right) + b_j \left(M_{ij} - \bar{M}_{.j} \right) + e_{ij}$$

Where $\bar{X_j}$ and $\bar{M}_{.j}$ represent The observed group mean was X and M, respectively. The upper-level equations are:

$$d_{Mj} = d_M a_B \bar{X}_{.j} + uM_j$$

$$d_{Yj} = d_Y + {c_B}' \bar{X}_{.j} b_B \bar{M}_{.j} + u_{Yj}$$

$$a_j = a_W + u_{aj}$$

$$b_j = b_W + u_{bj}$$

$$c'_j = c'_W + u_{c'j}$$

Using these subscripts, we can distinguish between the impacts within a group and the impacts between groups. The mean within-group indirect effect is

$$E(a_j b_j) = ab + \sigma_{(a_j b_j)}$$

Where b_j is the covariance between a_j and b_j. The between-group indirect effect is

$$E\left(a_B b_B\right) = ab + a_B b_B$$

5. Machine learning classifiers

5.1. Logistic regression model

The logistic regression model uses a sigmoid function to convert the continuous value output of the linear regression function into a definite value output. Any real-valued set of independent variables input into this sigmoid function, also known as the logistic function, is mapped into a value between 0 and 1.

Assume the independent input has the following characteristics:

$$X = \begin{bmatrix} x_{11} & \dots & x_{1m} \\ x_{21} & ldots & x_{2m} \\ \vdots & \ddots & \vdots \\ x_{n1} & \dots & x_{nm} \end{bmatrix}$$

Y is the dependent variable; it only has two possible values: 0 and 1.

$$Y = \begin{cases} 0 & \text{if Class of 1} \\ 1 & \text{if Class of 2} \end{cases}$$

Then, a multi-linear function will be applied to the input variables X.

$$z = \left(\sum_{i=1}^{n} w_i x_i\right) + b$$

Here x_i is the i^{th} observation of X, $w_i = [w_1, w_2, w_3, \cdots, w_m]$ is the weight, or Coefficient, and b is the bias term, also known as intercept. Simply put, this can be represented as the dot product of weight and bias.

$$z = w \cdot X + b$$

5.2. Logistic regression equations

The odd is the ratio of something occurring to something not happening. It differs from probability, as the probability is the ratio of something occurring to everything that could occur. So odd will be:

$$\frac{p(x)}{1 - p(x)} = e^z$$

Applying natural log on odd. Then, the log odd will be:

$$\begin{aligned}
\log\left[\frac{p(x)}{1 - p(x)}\right] &= z \\
\log\left[\frac{p(x)}{1 - p(x)}\right] &= w \cdot X + b \\
\frac{p(x)}{1 - p(x)} &= e^{w \cdot X + b} \quad \ldots \text{ Exponentiate both sides} \\
p(x) &= e^{w \cdot X + b} \cdot (1 - p(x)) \\
p(x) &= e^{w \cdot X + b} - e^{w \cdot X + b} \cdot p(x)\Big) \\
p(x) + e^{w \cdot X + b} \cdot p(x)\Big) &= e^{w \cdot X + b} \\
p(x)\left(1 + e^{w \cdot X + b}\right) &= e^{w \cdot X + b} \\
p(x) &= \frac{e^{w \cdot X + b}}{1 + e^{w \cdot X + b}}
\end{aligned}$$

Then, the final logistic regression equation will be:

$$p(X; b, w) = \frac{e^{wx+b}}{1 + e^{-X+b}} = \frac{1}{1 + e^{-wX+b}}$$

5.3. Likelihood function for logistic regression

The predicted probabilities will be:

- for $y = 1$, The predicted probabilities will be: $p(X; b, w) = p(x)$
- for $y = 0$, The predicted probabilities will be: $1 - p(X; b, w) = 1 - p(x)$

$$L(b, w) = \prod_{i=1}^{n} p\left(x_i\right)^{y_i} \left(1 - p\left(x_i\right)\right)^{1 - y_i}$$

Taking natural logs on both sides

$$\begin{aligned}\log(L(b,w)) &= \sum_{i=1}^{n} y_i \log p(x_i) + (1-y_i)\log(1-p(x_i)) \\ &= \sum_{i=1}^{n} y_i \log p(x_i) + \log(1-p(x_i)) - y_i \log(1-p(x_i)) \\ &= \sum_{i=1}^{n} \log(1-p(x_i)) + \sum_{i=1}^{n} y_i \log \frac{p(x_i)}{1-p(x_i} \\ &= \sum_{i=1}^{n} -\log 1 - e^{-(w\cdot x_i+b)} + \sum_{i=1}^{n} y_i (w\cdot x_i + b) \\ &= \sum_{i=1}^{n} -\log 1 + e^{w\cdot x_i+b} + \sum_{i=1}^{n} y_i (w\cdot x_i + b)\end{aligned}$$

5.4. SVC classifier

Given training vectors $x_i \in \mathbb{R}^p$, i = 1, …, n, in two classes, and a vector $y \in \{1,-1\}^n$, our goal is to find $w \in \mathbb{R}^p$ and $b \in \mathbb{R}$ such that the prediction given by sign $\left(w^T\phi(x)+b\right)$ is correct for most samples.

S.V.C. solves the following primal problem:

$$\begin{gathered}\min_{w,b,\zeta} \frac{1}{2}w^T w + C\sum_{i=1}^{n}\zeta_i \\ \text{subject to } y_i\left(w^T\phi(x_i)+b\right) \geq 1-\zeta_i \\ \zeta_i \geq 0, i=1,\ldots,n\end{gathered}$$

Intuitively, we're trying to maximize the margin by maximizing $\|w\|^2$ = minimizing while incurring a penalty when a sample is misclassified or within the margin boundary. Ideally, the value $y_i\left(w^T\phi(x_i)+b\right)$ would be ≥ 1 for all samples, which indicates a perfect prediction. But problems are usually not always perfectly separable with a hyperplane, so we allow some samples to be at a distance ζ_i from their correct margin boundary. The penalty term c controls the strength of this penalty and, as a result, acts as an inverse regularization problem with the primal is

$$\begin{gathered}\min_{\alpha} \frac{1}{2}\alpha^T Q\alpha - e^T\alpha \\ \text{subject to } y^T\alpha = 0 \\ 0 \leq \alpha_i \leq C, i=1,\ldots,n\end{gathered}$$

where e is the vector of all ones, and Q is an n by n positive semidefinite matrix, $Q_{ij} \equiv y_i y_j K(x_i,x_j)$, where $K(x_i,x_j) = \phi(x_i)^T\phi(x_j)$ is the kernel. The terms α_i are called the dual coefficients and are upper-bound by C. This dual representation highlights that training vectors implicitly map into a higher (maybe infinite) dimensional space by the function ϕ.

5.5. Linear SVC

The primal problem can be equivalently formulated as

$$\min_{w,b} \frac{1}{2}w^T w + C\sum_{i=1}^{n}\max\left(0, 1-y_i\left(w^T\phi(x_i)+b\right)\right)$$

Linear *nu*-SVC C directly optimizes this form where we use the hinge loss, but unlike the dual form, this one does not involve inner products between samples, so the well-known kernel trick is not applicable. This is why only the linear kernel is supported by Linear S.V.C. (ϕ is the identity function).

5.6. NuSVC

The ν-SVC formulation is a reparameterization SVC and, therefore, is mathematically equivalent.

We introduce a new parameter ν (instead of C), which controls the number of support vectors and margin errors: $\nu \in (0, 1]$ is an upper bound on the fraction of margin errors and a lower bound of the fraction of support vectors. A margin error corresponds to a sample on the wrong side of its margin boundary: it is either misclassified or correctly classified but does not lie beyond the margin.

Given training vectors $x_i \in \mathbb{R}^p$, i $= 1, \ldots,$ n, and a vector $y \in \mathbb{R}^n \varepsilon$-SVR solves the following primal problem:

$$\min_{w,b,\zeta,\zeta^*} \frac{1}{2} w^T w + C \sum_{i=1}^{n} (\zeta_i + \zeta_i^*)$$

$$\text{subject to } y_i - w^T \phi(x_i) - b \leq \varepsilon + \zeta_i$$

$$w^T \phi(x_i) + b - y_i \leq \varepsilon + \zeta_i^*$$

$$\zeta_i, \zeta_i^* \geq 0, i = 1, \ldots, n$$

Here, we penalize prediction as at least ε away from their target. These samples penalize the penalize by ζ_i or ζ_i^*, depending on whether their predictions lie above or below the ε tube.

where e is the vector of all ones, Q is an n by n positive semidefinite matrix, $Q_{ij} \equiv K(x_i, x_j) = \phi(x_i)^T \phi(x_j)$ is the kernel. Here, training vectors are implicitly mapped into a higher (maybe infinite) dimensional space by the function ϕ.

The prediction is:

$$\sum_{i \in SV} (\alpha_i - \alpha_i^*) K(x_i, x) + b$$

These parameters can be accessed through the attributes dual coefficient, which holds the difference $\alpha_i - \alpha_i^*$, support vectors, which hold the support vectors, and intercept, which has the independent term b.

5.7. TensorFlow

ANN of Keras model of the TensorFlow machine learning framework is mainly responsible for its flexibility and generality (Abadi, Isard, and Murray 2017). TensorFlow models are built from the ground up using familiar programming language techniques such as function composition. Deep neural network training and inference are among the machine learning applications that this method can make possible on heterogeneous distributed systems. Part of the core of some TensorFlow implementations is programming languages, like rewriting optimisations and optimisations. TensorFlow makes sense and is advantageous when viewed through the lens of programming languages. The main goal of the semantics is to provide a conceptual framework with an execution focus that can be used as a starting point for thinking about TensorFlow model behaviour. The main foundation of TensorFlow is dataflow graphs with modifiable states. The graph's semantics explain how diabetes and heart disease are predicted when other covariates are taken into account.

Tensor flow values include tensors, as expected, but also auxiliary values. Specifically, we assume:

- A set of values Tensors, which we may refer to as tensors.

- A constant GO that we use as a trigger.

- A constant EMPTY that we use to indicate not-yet-produced or already-consumed data. These are all disjoint. Correspondingly, we distinguish three kinds of edges:
- Tensor edges, which are used for conveying elements of Tensors;
- Variable edges, which are used for conveying elements of Vars; and Item Control edges, which are used only for G.O. signals.
- f for f a function in Tensors ^{k} → Tensors ^{l}, for some non-negative integers k and l;
- Var(x) for x in Vars;
- Assign-f for f a function in (Tensors × Tensors) → Tensors. When f is a function in Tensorsk →Tensorsl the operation f applies the function to the operation's k inputs and returns its l results.
- When x is a variable (an element of Vars), the operation Var(x) outputs x.
- The operation Read outputs the current value of a variable; this variable is an input to the operation.
- Finally, when f is a function in (Tensors × Tensors) → Tensors, the operation Assign-f has as inputs a variable x and a tensor v. it reads the current value of x, applies f to this value and v. and updates x to hold this result.

5.8. Accuracy, precision, recall and F-1 score

$$\begin{aligned} \text{Accuracy } &= \frac{TP+TN}{TP+TN+FP+FN} \\ \text{Precision } &= \frac{TP}{TP+FP} \\ \text{Recall } &= \frac{TP}{TP+FN} \end{aligned}$$

$$F1 = \frac{2 * \text{ Precision } * \text{ Recall}}{\text{Precision } + \text{ Recall}}$$

Note: Here, TP stands for true positive, TN is true negative, FP = False Positive and FN = False Negative.

5.9. Cross validation

For the k^{th} part, we fit the model to the other $K-1$ parts of the data and calculate the prediction error of the fitted model when predicting the k^{th} part of the data. We do this for $k = 1, 2, \ldots, K$ and combine the K estimates of prediction error.

Here are more details. Let $\kappa : \{1, \ldots, N\} \mapsto \{1, \ldots, K\}$ be an indexing function that indicates the partition to which the randomization i. Denote by $\hat{f}^{-k}(x)$ the fitted function, computed with the k^{th} part of the data removed. Then, the cross-validation estimate of prediction error is

$$\mathrm{CV}(\hat{f}) = \frac{1}{N} \sum_{i=1}^{N} L\left(y_i, \hat{f}^{-\kappa(i)}\left(x_i\right)\right).$$

Typical choices of K are 5 or 10 (see below). The case $K = N$ is known as leave-one-out cross-validation. In this case, $\kappa(i) = i$, and for the i^{th} observation, the fit is computed using all the data except the i^{th}.

6. Definitions

6.1. Odds ratio

A statistical metric called the odds ratio (O.R.) is used to express how strongly and in which direction two categorical variables are associated. It is frequently used in epidemiology, health research, and other statistical analysis domains. The odds ratio evaluates the likelihood that an event will occur in one group relative to the likelihood that it will occur in another.

The odds ratio can be computed using the following formula:

The odd's ratio can also be mathematically expressed as a probability:

$$\mathrm{OR} = \frac{p_1/\left(1-p_1\right)}{p_2/\left(1-p_2\right)}$$

Where, p_1 is the probability of the event occurring in Group 1, p_2 is the probability of the event occurring in Group 2.

6.2. Bootstraping

A lot of people use bootstrapping to figure out standard errors and confidence intervals for indirect effects in mediation analysis, as explained by Preacher and Hayes Preacher and Hayes (2008b). The method involves drawing repeated samples with replacements from the observed data, enabling more robust inference when sample sizes are limited.

6.3. Sobel test

Standard errors of a and b are represented, respectively, by s_a and s_b. The standard error of the indirect effect (s_{ab}) is given by Aroian, Mood, Graybill, and Boes, and Sobel (Preacher and Hayes 2004) as

$$s_{ab} = \sqrt{b^2 s_a^2 + a^2 s_b^2 + s_a^2 s_b^2}.$$

To conduct the test, ab is divided by s_{ab} to yield a critical ratio traditionally compared with the critical value from the standard normal distribution appropriate for a given alpha level. One of the assumptions necessary for the Sobel test is that the sample size is large, so the rough critical value for the two-tailed version of the test, assuming that the sampling distribution of ab is normal and that $\alpha = 0.05$, is ± 1.96. As the sample size becomes smaller, the Sobel test becomes less conservative. One variation of the Sobel test subtracts the last term of the standard error ($s_a^2 s_b^2$) rather than adding it Preacher and Hayes (2008a). Another version omits $s_a^2 s_b^2$ altogether because it is likely to be trivial (Baron & Kenny) LeBreton, Wu, and Bing (2009) describes a general procedure whereby more complicated indirect effects may be tested.

6.4. Sampling method and sample size

The study data is collected according to the National Family Health Survey (N.F.H.S.) guidelines. The published NFHS report (Iips 2017) provides a detailed description of the sampling design. This paper survey is based on a multi-stage cluster sampling design using the 2011 Census of India as a sampling framework to select primary sampling units representing rural and urban areas of Bhopal (Madhya Pradesh, India). The probability proportional to size (PPS) sampling method is used to pick an equal number of units from both urban and rural areas so that the sample is representative of a wide range of healthcare settings and socioeconomic backgrounds. Details the process of selecting healthcare facilities within these PSUs, aiming for a mix of 2-3 hospitals or clinics per PSU that cater to diabetes, heart disease, and obesity. This ensures that the study encompasses various healthcare providers, including public and private institutions, to represent different socioeconomic groups adequately. Focusses on the systematic selection of patients within these facilities. Eligibility criteria are applied to identify individuals over 18 years of age diagnosed with diabetes and heart disease who also have a Body Mass Index (BMI) $\geq$ 25, indicating obesity. The criteria exclude individuals below 18 years, those severely ill, or those with gestational diabetes. The aim is to select a total of 421 participants, which was determined to be the necessary sample size based on the 95% CI, 4.78% margin of error, 1.5 design effects, and >10% non-response rate., with consideration for oversampling to account for potential non-responses or ineligible participants, targeting an initial sample size of approximately 460-470 individuals. The methodology highlights the significance of implementing a systematic sampling approach to guarantee that the study encompasses a wide range of the population impacted by these conditions in various geographic regions and healthcare environments. According to the NFHS-5 factsheet for India, 14% of adults aged 18 and above had blood sugar levels higher than 140 mg/dl, which is considered high.

6.5. Data collection method

The International Institute for Population Sciences conducted the study and used the men's and women's files from the National Family Health Survey rounds 4 and 5, respectively, from 2019 to 2021, for the current cross-sectional study. The research survey is carried out in accordance with the guidelines provided by N.F.H.S.

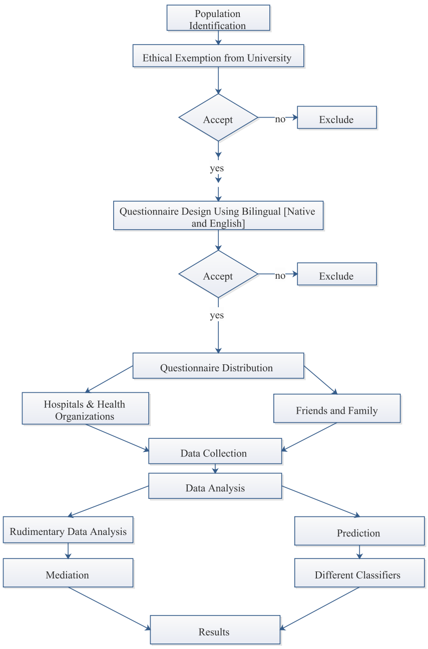
Population Identification
Ethical Exemption from University
Accept
no
Exclude
yes
Questionnaire Design Using Bilingual [Native and English]
Accept
no
Exclude
yes
Questionnaire Distribution
Hospitals & Health Organizations
Friends and Family
Data Collection
Data Analysis
Rudimentary Data Analysis
Prediction
Mediation
Different Classifiers
Results

Figure 2: Overview of data collection

6.6. Criteria for inclusion and exclusion

Type-2 diabetes patients aged 18 years and older are included in the study (Al-Rubeaan 2015). Patients under 18 have not been included who were found seriously ill and having gestational diabetes.

6.7. Quantitative tools

The surveys were conducted verbally and through questionnaires, obtaining quantitative information. Questionnaires were prepared based on NFHS-5 guidelines, containing all the relevant questions regarding the cause of diabetes and associated factors.

6.8. Behavior and lifestyle

We collected the respondent's behavioural and lifestyle patterns by asking how often they drink, consume tobacco, gutkha, and smoke. Respondents were divided into high- and low-risk categories based on their responses. (Gupte, Mandal, and Chatterjee 2023).

6.9. BMI and blood sugar levels

Medical reports of respondents were used to collect data on sugar levels (fasting and after-meal) (Dimitriadis, Mitrou, Lambadiari, Boutati, Maratou, Koukkou, Tzanela, Thalassinos, and Raptis 2006). Fasting sugar levels should be 100 grams/dl, and post-meal sugar levels should be 140 grams/dl (Bozkaya, Ozgu, and Karaca 2010). The Ministry of Ayurveda, Yoga, and Naturopathy, Unani, Siddha and Homoeopathy (A.Y.U.S.H.), Government of India, indicates that sugar levels above these values indicate diabetic conditions (Mathur, Leburu, and Kulothungan 2022). Based on the medical reports of people with diabetes, types of morbidities were also determined (Panda, Ratha, and Rao 2017). Furthermore, data on categories of morbidities was gathered from diabetic patients' medical records. The respondent's BMI was calculated using their height and weight. An underweight BMI of less than 18.5 indicates underweight, an average BMI of 18.5–25 indicates normal weight, and a BMI of 25 or higher indicates overweight or obesity (Stommel and Schoenborn 2009).

6.10. Sleep pattern

Adults are advised to have 7-8 hours of restorative sleep per night. Still, according to national statistics, the frequency of inadequate sleep has grown over the previous ten years across all adult age and sex categories (Kocevska, Lysen, Dotinga, Koopman-Verhoeff, Luijk, Antypa, Biermasz, Blokstra, Brug, Burk *et al.* 2021). Controlling several physiological processes that are related to metabolism requires rest. Because of this, the studies suggest a link between sleep habits and the risk of developing diabetes. Diabetes risk factors include insufficient sleep length, lab sleep restriction, poor sleep quality, and sleep disorders such as insomnia and sleep apnea (Grandner, Seixas, Shetty, and Shenoy 2016). Little sleep and daytime naps increase the risk of developing diabetes. The frequency of naps may alter the link between sleep duration and diabetes (Xu, Song, Hollenbeck, Blair, Schatzkin, and Chen 2010).

6.11. Description of variables

A finger-stick blood specimen was collected using a freestyle optimum glucometer for the NFHS-5. Based on random glucose levels, NFHS-5 classified men and women as 'high' for levels 141–160 mg/dl and 'very high' for levels $>$ 160 mg/dl (Pengpid and Peltzer 2019). According to some researchers, the random cut-off point for 2-h plasma glucose greater than 200 mg./dl (11.1 mmol/l) is equivalent to 140 mg./dl (7.7 mmol/l) for capillary blood glucose (Kuzuya, Nakagawa, Satoh, Kanazawa, Iwamoto, Kobayashi, Nanjo, Sasaki, Seino, Ito *et al.* 2002). This study follows the exact definition of excessive blood glucose levels as others, having the random glucose level above 140 mg/dl as above average (Ghosh, Dhillon, and Agrawal 2020).

7. Statistical analysis and interpretations

7.1. Descriptive statistics

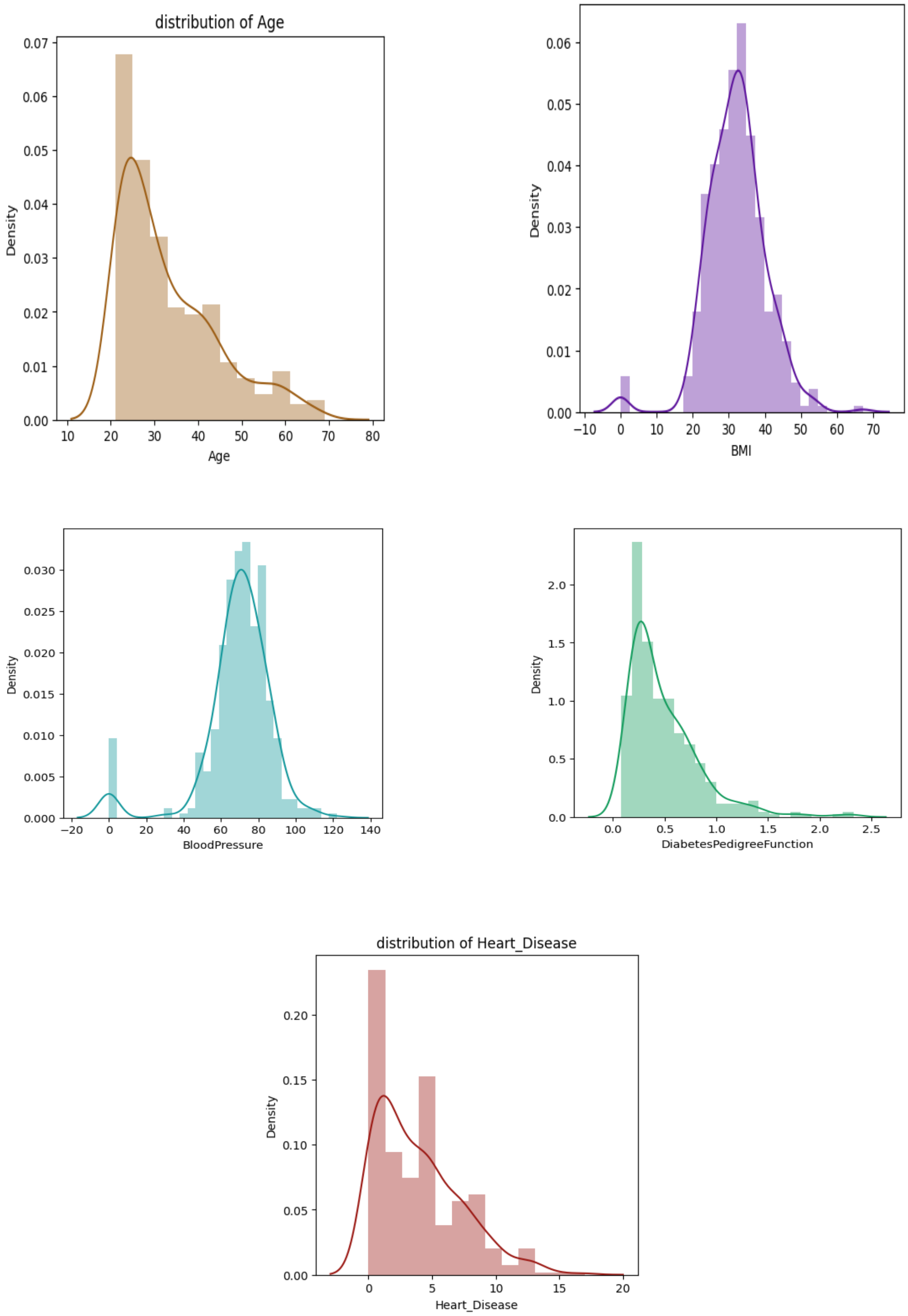

Figure 3: Descriptive measure of the variables

Figure 3 depicts the structure of the major variables used for model. For data analysis, Hayes Macro Process is used for mediation analysis with R software Hayes (2012). Mediation analysis has been used to assess the association of BMI with diabetes and heart disease.

In the present study, the dependent variable (Y) is Heart Disease, the Independent variable (X) is diabetes, and the mediator variable (M) is BMI. For greater accuracy of the model, we have included some other covariates defined Age, sleep time, smoking, and alcohol drinking; the sample size is 421; the random seed is 756890; and the outcome variable is BMI

Table 1: Model summary of mediation model

R	R-square	MSE	F	df1	df2	p-value
0.7124	0.5076	16.3039	71.1252	6.0	414.0	0.00

Table 1 show that the value of R^2 is 0.5076, i.e., 50.76%, which reveals the variability observed by the mediator variable BMI. It depicts that the model supports the mediation between the selected variables for the study, which the regression model explains.

Table 2: Summary of mediation analysis with outcome variable BMI

	Coefficient	se	t	p-value	LLCI	ULCI
constant	24.4203	1.2144	20.1083	0.000	22.0330	26.8075
Diabetic	7.7807	0.3961	19.6439	0.000	7.0021	8.5593
Age	-0.0409	0.0174	-2.3501	0.0192	-0.0752	-0.0067
Sex	1.4962	0.4004	3.7366	0.0002	0.7091	2.2833
Sleep time	-0.0023	0.3951	-0.0057	0.9954	-0.7789	0.7744
Smoking	-0.8203	0.4145	-1.9789	0.0485	-1.6352	-0.0055
Alcohol. Drinking	-0.1973	0.8335	-0.2367	0.8130	-1.8357	1.4412

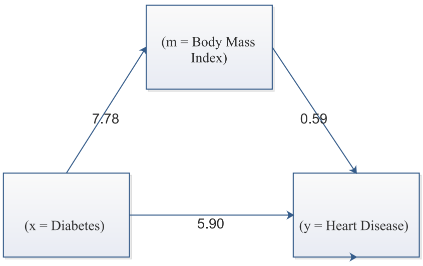

Figure 4: Relations between dependent and independent variables with mediator

In Figure 4 and Table 2, the path (direct effect) from diabetic to BMI is positive and statistically significant (b = 7.7807, se = 0.3961, O.R. = 2393.94, 95% confidence interval (7.0021, 8.5593), p = 0.00). The path (direct effect) of Age to BMI is negative and statistically significant (b = -0.0409, S.E. = 0.0174, O.R. = 0.959, 95% confidence interval (-0.0752, -0.0067), p = 0.0192). The path (direct effect) from Sex to BMI is positive and statistically significant (b = 1.4962, se = 0.4004, O.R. = 4.464, 95% confidence interval (0.7091, 2.2833), p = 0.0002). The path (direct effect) from sleep time to BMI is negative and statistically not significant (b = -0.0023, se = 0.3951, O.R. = 0.997, 95% confidence interval (-0.7789, 0.7744), p = 0.9954). The path (direct effect) from smoking to BMI is negative and statistically significant (b = -0.8203, se = 0.4145, O.R. = 0.4402, 95% confidence interval (-1.6352, -0.0055), p = 0.0485). The path (direct effect) from Alcohol drinking to BMI is negative and statistically non-significant (b = -0.1973, se = 0.8335, O.R. = 0.8209, 95% confidence interval (-1.8357, 1.4412), p = 0.8130).

Table 3: Model summary of Table 4

$-2LL$	ModelLL	df	p	McFadden	CoxSnell	Nagelkrk
89.99	493.09	7.0000	0.0000	.8457	0.6900	0.9204

In Table 4, The path (direct effect) from diabetes to BMI is positive and statistically significant (b = 5.9010, S.E. = 0.8240, O.R. = 365.40, 95% confidence interval (4.2860, 7.5161), p =

Table 4: Summary of mediation analysis with outcome variable heart disease

	Coefficient	se	t	p-value	LLCI	ULCI
Constant	19.2658	3.7750	-5.1035	0.0000	-26.6647	-11.8668
Diabetic	5.9010	0.8240	7.1614	0.0000	4.2860	7.5161
BMI	0.5942	0.1209	4.9140	0.0000	0.3572	0.8312
Age	-0.0037	0.0266	-0.1389	0.8895	-0.0558	0.0485
Sex	0.5563	0.6459	0.8612	0.3891	-0.7097	1.8223
Sleep Time	-1.0829	0.6324	-1.7124	0.0868	-2.3223	0.1566
Smoking	2.2117	0.8083	2.7361	0.0062	0.6274	3.7960
Alcohol Drinking	0.2822	1.2611	0.2238	0.8229	-2.1896	2.7539

0.00) indicating that persons who have diabetes they have more chances for heart disease those having high BMI The direct effect of Age on heart disease is negative and statistically non-significant (b = -0.0037 S.E. = 0.0266, O.R. = 0.9963, 95% confidence interval (-0.0558, 0.0485), p = 0.8895) indicating respondent Age is negatively correlated with Heart disease. The direct effect of SeSexn heart disease is statistically non-significant (b = 0.5563 S.E. = 0.6459, O.R. = 1.744, 95% confidence interval (-0.7097, 1.8223), p = 0.3891) indicating respondent Age is negatively correlated with Heart disease. The direct effect of Sleep time on heart disease is statistically non-significant (b = -1.0829, S.E. = 0.6324, O.R. = 0.3386, 95% confidence interval (-2.3223, 0.1566), p = 0.0868) indicating the respondent's sleep time is not correlated with Heart disease. The direct effect of smoking on heart disease is positive and statistically significant (b = 2.2117, S.E. = 0.8083, O.R. = 9.131, 95% confidence interval (0.6274, 3.7960), p = 0.0062) indicating respondent Age is correlated with Heart disease. The direct effect of AlcohoAlcoholing on heart disease is statistically non-significant (b = 0.2822 S.E. = 1.2611, O.R. = 1.744, 95% confidence interval (-2.1896, 2.7539), p = 0.8229) indicating respondent Age is negatively correlated with Heart disease.

Table 5: Direct effect of diabetes on heart disease

Effect	se	Z	p	LLCI	ULCI
5.9010	0.8240	7.1614	0.000	4.2860	7.5161

Table 6: Indirect effect(s) of diabetes on heart disease

	Effect	Boot SE	Boot LLCI	BootULCI
BMI	4.6231	11.8542	3.1333	9.6556

Tables 5 and 6 depict the result of testing the indirect effect using non-parametric bootstrapping. In this case, the Indirect impact is 4.6231, which is statistically significant at a 95% confidence interval (3.133, 9.6556). The indirect effect of Diabetes on Heart disease is also substantial because BootLLCI and BootULCI are positive, and zero does not lie between the two limits. Hence, a considerable mediation of BMI on diabetes and heart disease exists.

8. Machine learning classifiers

In this analysis, we have used four machine-learning algorithms to predict diabetes and heart disease. In the given table 7, it is depicted that the Tensorflow gives the highest accuracy of 99%, the F1-score is 99%, and the K-fold cross-validation is 98%, which shows the perfect prediction of the mentioned disease. The other three classifiers, Linear S.V.C., NuSVC, and Logistic regression, have less accuracy and precision than the Tensorflow classifier.

Table 7: Classification report of machine learning algorithms

Classifiers	Accuracy	Precision	F-1 Score	Cross Validation
TensorFlow	0.99	0.97	0.99	0.98
Linear SVC Classifier	0.98	0.97	0.96	0.95
NuSVC Classifier	0.96	0.95	0.98	0.97
Logistic Regression Classifier	0.97	0.96	0.98	0.99

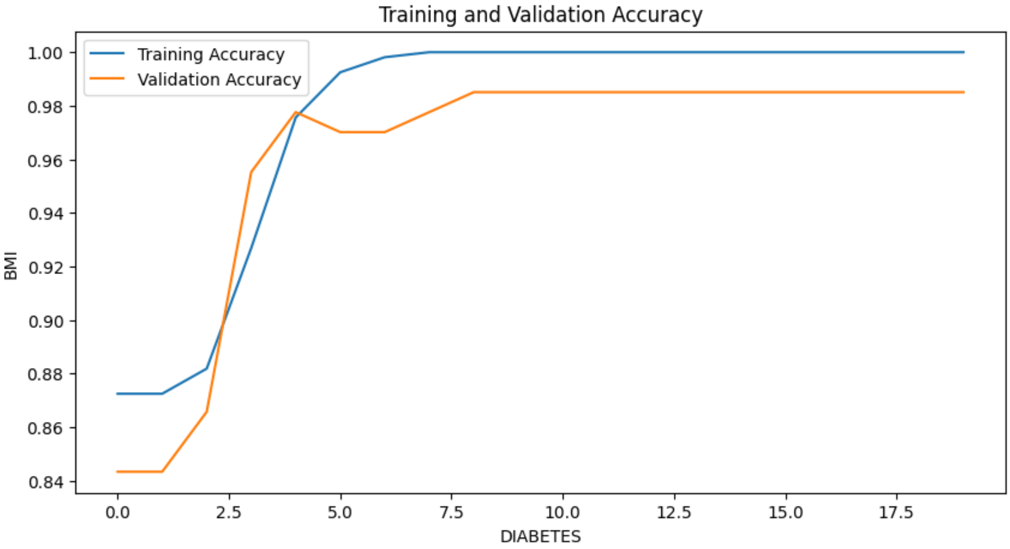

Figure 5: TensorFlow

9. Comparison of machine learning classifiers with mediation model

Our study employed several machine learning classifiers, including TensorFlow S.V.C., NuSVC and Logistic Regression classifiers, to predict the risk of diabetes and heart disease. Table 7 and Figure 5 display the performance metrics, including accuracy, precision, recall, F1-score, and k-fold cross-validation, for each classifier. Key factors impacting the risk of diabetes and heart disease were identified via feature importance analysis. Mediation analysis was done to investigate potential mediating factors in the link between predictors and outcomes. There were notable collateral effects [BMI partially caused diabetes and heart disease]. This shows that specific variables may mediate the onset of diabetes and heart disease. The risk prediction model's intricate relationships are comprehensively understood by utilizing it to reveal each path's strength and significance. Our machine learning models performed exceptionally well in estimating the risk of heart disease and diabetes.

10. Results and discussion

According to the study's findings, BMI also acts as a bridge between diabetes and heart disease. Raising BMI levels can achieve it, which indirectly affects diabetes and heart disease. The current investigation aimed to determine whether BMI mediated the link between diabetes and heart disease. One of the primary elements of non-communicable diseases is diabetes (NCDs). A rise in diabetes cases has been associated with several risk variables, classified as modifiable and non-modifiable risk factors. Modifiable risk factors include obesity, sedentary behaviour, poor diet, stress, alcohol usage, and viral infectionsRamachandran, Ma, and Snehalatha (2010). Modifiable factors include obesity, sedentary behaviour, poor diet, anxiety, smoking, drinking, and viral disease. Family history, genetic predisposition to the illness, and socioeconomic level are risk factors that cannot be alteredKinra, Bowen, Lyngdoh, Prabhakaran, Reddy, Ramakrishnan, Gupta, Bharathi, Vaz, Kurpad *et al.* (2010). The findings of this study indicated that BMI has a significant direct effect on the risk of diabetes and heart disease, as well as an indirect effect mediated by blood pressure, fasting blood sugar, and cholesterol levels. This suggests that interventions to reduce BMI could potentially decrease the prevalence of diabetes and heart disease. The risk prediction model's intricate relationships are comprehensively understood using each path's strength and significance. Our machine learning models performed exceptionally well in estimating the risk of heart disease and diabetes.

11. Conclusion

The outcomes of this study provide valuable insights into the complex relationship between BMI, diabetes, and heart disease and highlight the need for further research and targeted public health interventions to address this critical public health issue. Our knowledge of the intricate relationships underlying the risk of diabetes and heart disease could be significantly advanced by integrating machine learning classifiers with mediation analysis. This novel method helps us create more accurate predictive models specific to the Indian environment and makes it easier to identify essential mediators. To predict the risk of diabetes and heart disease, our study used several machine learning classifiers, such as TensoFlow S.V.C., NuSVC, and logistic regression classifiers. The performance measures for each classifier, such as accuracy, precision, recall, F1-score, and k-fold cross-validation, are shown in Table 7 and Figure 10. According to this study, gaining weight is the primary cause of the disease that people with diabetes experience. Other factors contributing to diabetes include lack of sleep, inconsistent eating patterns, and improper chewing of each bite. Alcohol consumption and smoking both significantly affect diabetes. People with diabetes also develop heart disease very quickly due to this. Because of their increased dietary intake of calories, fat, and sugar, most diabetics and people with heart disease have obesity and high blood sugar levels. The prevalence of overweight or obese individuals and diabetes has increased due to these risk factors. The prevalence of various comorbid illnesses (high blood sugar levels, incorrect food intake, chewing of food, lack of exercise, physical sickness, mental illness, cardiovascular disease, increased blood pressure, smoking, and alcohol intake) is higher among individuals with diabetes: increased sedentary behaviour, drug usage, inactivity, and unconsciousness. Comorbidities were inevitable in diabetes treatment. As a result, spreading awareness of diabetes prevention and treatment through illness control programs, health education, educated community and peer medical professionals, and community-based programs is essential Doherty (2015).

12. Future work

Building on the groundwork established by this work, future research avenues should be able to further our comprehension of the interactions among obesity, diabetes, and heart disease by applying cutting-edge machine-learning techniques. Future research can create more thorough models and successful interventions by combining longitudinal data, more variables, and various population types. He will ultimately improve health outcomes and direct public health initiatives.

Longitudinal studies Conducting longitudinal studies to observe trends and changes in BMI, diabetes, and heart disease over time provides a more comprehensive understanding of the relationships and potential causal pathways.

Causal inference methods Implementing advanced causal inference methods, such as instrumental variable analysis or propensity score matching, to strengthen the causal interpretation of the relationships observed in the mediation analysis.

Exploration of specific mediators Delving deeper into specific mediators identified in the mediation analysis, such as exploring the role of lifestyle factors, genetic predispositions, or socioeconomic variables in mediating the relationship between BMI, diabetes and heart disease

Regional variations Investigating regional variations within India to understand how cultural, environmental, and regional designers may influence the observed relationships, providing insights for more targeted interventions.

Intervention studies Designing and implementing intervention studies based on the findings to develop effective strategies to mitigate the impact of high BMI on diabetes and heart disease outcomes in the Indian population.

Acknowledgements

The author thanks all participants who responded to the questionnaire.

References

Abadi M, Isard M, Murray DG (2017). “A Computational Model for TensorFlow: An Introduction.” In *Proceedings of the 1st ACM SIGPLAN International Workshop on Machine Learning and Programming Languages*, pp. 1–7.

Abbott RD (1985). “Logistic Regression in Survival Analysis.” *American Journal of Epidemiology*, **121**(3), 465–471.

Al-Rubeaan K (2015). “National Surveillance for Type 1, Type 2 Diabetes and Prediabetes among Children and Adolescents: A Population-based Study (SAUDI-DM).” *Journal of Epidemiology and Community Health*, **69**(11), 1045–1051.

Bozkaya G, Ozgu E, Karaca B (2010). “The Association between Estimated Average Glucose Levels and Fasting Plasma Glucose Levels.” *Clinics*, **65**(11), 1077–1080.

Collaboration PS, *et al.* (2009). “Body-mass Index and Cause-specific Mortality in 900 000 Adults: Collaborative Analyses of 57 Prospective Studies.” *The Lancet*, **373**(9669), 1083–1096.

de Lucia C, Metzinger L, Wallner M (2023). *Diabetes and Heart Failure: Basic, Translational, and Clinical Research.* Frontiers Media SA.

Deepa M, Bhansali A, Anjana RM, Pradeepa R, Joshi SR, Joshi PP, Dhandhania VK, Rao PV, Subashini R, Unnikrishnan R, *et al.* (2014). “Knowledge and Awareness of Diabetes in Urban and Rural India: The Indian Council of Medical Research India Diabetes Study (Phase I): Indian Council of Medical Research India Diabetes 4.” *Indian Journal of Endocrinology and Metabolism*, **18**(3), 379.

Dimitriadis G, Mitrou P, Lambadiari V, Boutati E, Maratou E, Koukkou E, Tzanela M, Thalassinos N, Raptis SA (2006). “Glucose and Lipid Fluxes in the Adipose Tissue after Meal Ingestion in Hyperthyroidism.” *The Journal of Clinical Endocrinology & Metabolism*, **91**(3), 1112–1118.

Doherty AM (2015). “Psychiatric Aspects of Diabetes Mellitus.” *BJPsych Advances*, **21**(6), 407–416.

Dudina A, Cooney MT, Bacquer DD, Backer GD, Ducimetière P, Jousilahti P, Keil U, Menotti A, Njølstad I, Oganov R, *et al.* (2011). "Relationships between Body Mass Index, Cardiovascular Mortality, and Risk Factors: A Report from the SCORE Investigators." *European Journal of Cardiovascular Prevention & Rehabilitation*, **18**(5), 731–742.

Farooqi A, Khunti K, Abner S, Gillies C, Morriss R, Seidu S (2019). "Comorbid Depression and Risk of Cardiac Events and Cardiac Mortality in People with Diabetes: A Systematic Review and Meta-analysis." *Diabetes Research and Clinical Practice*, **156**, 107816.

Ghosh K, Dhillon P, Agrawal G (2020). "Prevalence and Detecting Spatial Clustering of Diabetes at the District Level in India." *Journal of Public Health*, **28**, 535–545.

Gierach M, Gierach J, Ewertowska M, Arndt A, Junik R (2014). "Correlation between Body Mass Index and Waist Circumference in Patients with Metabolic Syndrome." *International Scholarly Research Notices*, **2014**.

Grandner MA, Seixas A, Shetty S, Shenoy S (2016). "Sleep Duration and Diabetes Risk: Population Trends and Potential Mechanisms." *Current Diabetes Reports*, **16**, 1–14.

Gupta A, Kumar R, Arora HS, Raman B (2022). "C-CADZ: Computational Intelligence System for Coronary Artery Disease Detection Using Z-Alizadeh Sani Dataset." *Applied Intelligence*, **52**(3), 2436–2464.

Gupte HA, Mandal G, Chatterjee N (2023). "Sociodemographic Factors, Attitudes, and Tobacco Use among Adolescent Areca-Nut Users in Mumbai, India." *Indian Journal of Community Medicine: Official Publication of Indian Association of Preventive & Social Medicine*, **48**(1), 183.

Hayes AF (2012). "PROCESS: A Versatile Computational Tool for Observed Variable Mediation, Moderation, and Conditional Process Modeling."

Hruby A, Hu FB (2015). "The Epidemiology of Obesity: A Big Picture." *Pharmacoeconomics*, **33**, 673–689.

Iips ICF (2017). "India National Family Health Survey NFHS-4 2015–16." *Mumbai: IIPS and ICF*, pp. 1255–9.

Joachims T, Joachims T (2002). "Support Vector Machines." *Learning to Classify Text Using Support Vector Machines*, pp. 35–44.

Joshi SR (2015). "Diabetes Care in India." *Annals of Global health*, **81**(6), 830–838.

Kaveeshwar SA, Cornwall J (2014). "The Current State of Diabetes Mellitus in India." *The Australasian Medical Journal*, **7**(1), 45.

Ketabchi S, Moosaei H, Razzaghi M, Pardalos PM (2019). "An Improvement on Parametric ν-support Vector Algorithm for Classification." *Annals of Operations Research*, **276**(1), 155–168.

Kinra S, Bowen LJ, Lyngdoh T, Prabhakaran D, Reddy KS, Ramakrishnan L, Gupta R, Bharathi AV, Vaz M, Kurpad AV, *et al.* (2010). "Sociodemographic Patterning of Non-communicable Disease Risk Factors in Rural India: A Cross Sectional Study." *BMJ*, **341**.

Kleinbaum DG, Dietz K, Gail M, Klein M, Klein M (2002). *Logistic Regression.* Springer.

Kocevska D, Lysen TS, Dotinga A, Koopman-Verhoeff ME, Luijk MP, Antypa N, Biermasz NR, Blokstra A, Brug J, Burk WJ, *et al.* (2021). "Sleep Characteristics across the Lifespan in 1.1 Million People from the Netherlands, United Kingdom and United States: A Systematic Review and Meta-analysis." *Nature Human Behaviour*, **5**(1), 113–122.

Kuzuya T, Nakagawa S, Satoh J, Kanazawa Y, Iwamoto Y, Kobayashi M, Nanjo K, Sasaki A, Seino Y, Ito C, *et al.* (2002). "Report of the Committee on the Classification and Diagnostic Criteria of Diabetes Mellitus." *Diabetes Research and Clinical Practice*, **55**(1), 65–85.

Lakshminarayan V, Tejaswi S (2014). *Diabetes: India's Invisible Enemy.* Vani Prakashan.

LeBreton JM, Wu J, Bing MN (2009). "The Truth(s) on Testing for Mediation in the Social and Organizational Sciences." *Statistical and Methodological Myths and Urban Legends: Doctrine, Verity and Fable in the Organizational and Social Sciences*, **78**, 107–141.

Longo M, Zatterale F, Naderi J, Parrillo L, Formisano P, Raciti GA, Beguinot F, Miele C (2019). "Adipose Tissue Dysfunction as Determinant of Obesity-associated Metabolic Complications." *International Journal of Molecular Sciences*, **20**(9), 2358.

Lutz B, Zwygart S, Thomann B, Stucki D, Burla JB (2022). "The Relationship between Common Data-based Indicators and the Welfare of Swiss Dairy Herds." *Frontiers in Veterinary Science*, **9**, 991363.

Maiti S, Akhtar S, Upadhyay AK, Mohanty SK (2023). "Socioeconomic Inequality in Awareness, Treatment and Control of Diabetes among Adults in India: Evidence from National Family Health Survey of India (NFHS), 2019–2021." *Scientific Reports*, **13**(1), 2971.

Mathur P, Leburu S, Kulothungan V (2022). "Prevalence, Awareness, Treatment and Control of Diabetes in India from the Countrywide National NCD Monitoring Survey." *Frontiers in Public Health*, **10**, 748157.

Murugesan N, Snehalatha C, Shobhana R, Roglic G, Ramachandran A (2007). "Awareness about Diabetes and Its Complications in the General and Diabetic Population in a City in Southern India." *Diabetes Research and Clinical Practice*, **77**(3), 433–437.

Nguyen TQ, Ogburn EL, Schmid I, Sarker EB, Greifer N, Koning IM, Stuart EA (2023). "Causal Mediation Analysis: From Simple to More Robust Strategies for Estimation of Marginal Natural (in) Direct Effects." *Statistics Surveys*, **17**, 1.

Panda AK, Ratha KK, Rao MM (2017). "Efficacy of Ayurveda Formulation Ayush-82 (IME-9) in Newly Diagnosed Type 2 Diabetics: Retrospective Analysis of Individual Data." *Journal of Traditional Medicine and Clinical Naturopathy*, **6**(250), 2.

Pengpid S, Peltzer K (2019). "Prevalence and Correlates of Underweight and Overweight/Obesity among Women in India: Results from the National Family Health Survey 2015–2016." *Diabetes, Metabolic Syndrome and Obesity: Targets and Therapy*, pp. 647–653.

Poirier P, Giles TD, Bray GA, Hong Y, Stern JS, Pi-Sunyer FX, Eckel RH (2006). "Obesity and Cardiovascular Disease: Pathophysiology, Evaluation, and Effect of Weight Loss: An Update of the 1997 American Heart Association Scientific Statement on Obesity and Heart Disease from the Obesity Committee of the Council on Nutrition, Physical Activity, and Metabolism." *Circulation*, **113**(6), 898–918.

Preacher KJ, Hayes AF (2004). "SPSS and SAS Procedures for Estimating Indirect Effects in Simple Mediation Models." *Behavior Research Methods, Instruments, & Computers*, **36**, 717–731.

Preacher KJ, Hayes AF (2008a). *Assessing Mediation in Communication Research.* The Sage Sourcebook of Advanced Data Analysis Methods for Communication ....

Preacher KJ, Hayes AF (2008b). "Asymptotic and Resampling Strategies for Assessing and Comparing Indirect Effects in Multiple Mediator Models." *Behavior Research Methods*, **40**(3), 879–891.

Ramachandran A, Ma RCW, Snehalatha C (2010). "Diabetes in Asia." *The Lancet*, **375**(9712), 408–418.

Rani PK, Raman R, Subramani S, Perumal G, Kumaramanickavel G, Sharma T (2008). "Knowledge of Diabetes and Diabetic Retinopathy among Rural Populations in India, and the Influence of Knowledge of Diabetic Retinopathy on Attitude and Practice." *Rural and Remote Health*, **8**(3), 1–9.

Rathmann W, Giani G (2004). "Global Prevalence of Diabetes: Estimates for the Year 2000 and Projections for 2030: Response to Wild et al." *Diabetes Care*, **27**(10), 2568–2569.

Seidell JC, Hautvast JGAJ, Deurenberg P (1989). "Overweight: Fat Distribution and Health Risks. Epidemiological Observations." *Transfusion Medicine and Hemotherapy*, **16**(6), 276–281.

Shah NR, Braverman ER (2012). "Measuring Adiposity in Patients: The Utility of Body Mass Index (BMI), Percent Body Fat, and Leptin." *PloS One*, **7**(4), e33308.

Shetty R, Jena B, Kadithi A (2013). "Can Social Scientists Be the Change Agents for Diabetes Prevention? Diabetes-related Knowledge, Attitude, and Practice among Social Scientists." *Journal of Social Health and Diabetes*, **1**(01), 032–036.

Stommel M, Schoenborn CA (2009). "Accuracy and Usefulness of BMI Measures Based on Self-reported Weight and Height: Findings from the NHANES & NHIS 2001-2006." *BMC Public Health*, **9**, 1–10.

Tantengco OAG (2022). "Decreased Global Online Interest in Obesity from 2004 to 2021: An Infodemiology Study." *Obesity Medicine*, **30**, 100389.

Wang P, Wang Q, Yang B, Zhao S, Kuang H (2015). "The Progress of Metabolomics Study in Traditional Chinese Medicine Research." *The American Journal of Chinese Medicine*, **43**(07), 1281–1310.

Wells JCK, Pomeroy E, Walimbe SR, Popkin BM, Yajnik CS (2016). "The Elevated Susceptibility to Diabetes in India: An Evolutionary Perspective." *Frontiers in Public Health*, **4**, 145.

Wilson JR, Lorenz KA, Wilson JR, Lorenz KA (2015). "Standard Binary Logistic Regression Model." *Modeling Binary Correlated Responses Using SAS, SPSS and R*, pp. 25–54.

Xu Q, Song Y, Hollenbeck A, Blair A, Schatzkin A, Chen H (2010). "Day Napping and Short Night Sleeping Are Associated with Higher Risk of Diabetes in Older Adults." *Diabetes Care*, **33**(1), 78–83.

Affiliation:

Ajay Verma
Mathematics Division,
School of Advanced Sciences and Languages
VIT Bhopal University
Sehore M.P. India, 466114
E-mail: ajai.varma1729@gmail.com

Manisha Jain
Mathematics Division,
School of Advanced Sciences and Languages
VIT Bhopal University
Sehore M.P. India, 466114
E-mail: mujain31@gmail.com

Austrian Journal of Statistics http://www.ajs.or.at/
published by the Austrian Society of Statistics http://www.osg.or.at/
Volume 53
2024
Submitted: 2024-04-27
Accepted: 2024-07-05