Modeling Health Risk Probabilities of Obesity and Hypertension Using Machine Learning

Ramón Moreno Ríos
Master of Engineering in Computer Engineering
Advisor: Nelliud Torres Batista, DBA
Polytechnic University of Puerto Rico
Graduate Project EXPO, May 2026

***Abstract* —** *This study investigates the use of machine learning techniques to model health risk probabilities using data from the National Health and Nutrition Examination Survey 2017–2018. The objective is to evaluate the predictive performance of multiple models across health outcomes, including hypertension, stroke, kidney disease, and sleep disorders, with a focus on the relationship between obesity and hypertension. Variables such as body mass index, waist circumference, age, gender, and blood pressure measurements were used to develop predictive models. The results indicate that hypertension prediction achieved the highest performance, while stroke and kidney disease showed moderate accuracy. Sleep-related outcomes showed reduced predictive power due to limited feature representation. Additionally, the analysis confirmed a strong relationship between obesity and hypertension, as individuals with higher body mass index exhibited increased blood pressure levels. Overall, the findings highlight the potential of machine learning in health risk prediction while emphasizing the importance of data quality and feature selection.*

***Keywords* —** *Hypertension, Machine Learning, Obesity, Probability.*

INTRODUCTION

High blood pressure and obesity are increasingly prevalent chronic health issues that increase the risk of cardiovascular and metabolic complications [1], [2], [3]. The body mass index is used in the United States to categorize the severity of obesity, which is still very common in adults [1]. High blood pressure, also known as hypertension, is frequently referred to as a "silent" illness since many people may not have symptoms, but it is connected to significant consequences if it is not diagnosed and treated [2], [3]. Probability-based risk modeling may offer more informative predictions than single-threshold screening alone, since these disorders often coexist and risk depends on several interacting variables (e.g., body mass index, repeated blood pressure measurements, demographic features, and laboratory indicators) [2], [3].

Large-scale population health studies enable this type of modeling using standardized measurements collected across diverse participants. The National Health and Nutrition Examination Survey combines interviews with physical examinations and laboratory testing, providing measured body mass index, repeated blood pressure readings, and outcome-related variables, all documented in official codebooks for each survey component [4], [5], [6]. Using these documented variables, the present work develops machine learning models, including a neural network, to estimate outcome probabilities related to obesity and hypertension, and it also examines the relationship between body mass index and hypertension using descriptive visual analysis.

BACKGROUND

Obesity is a major global public health challenge, affecting millions across all age groups. It is defined as excessive body fat that poses health risks, typically measured by body mass index (BMI). The World Health Organization reports a dramatic increase in obesity rates over recent decades, which has contributed to higher rates of chronic diseases such as cardiovascular disease, diabetes, and other metabolic conditions [7]. In the United States, national surveillance data show that adult obesity remains prevalent, underscoring its significant impact on long-term health [8].

Hypertension, or high blood pressure, is a serious health condition strongly linked to obesity. It is a leading risk factor for cardiovascular diseases such as heart attacks and stroke. Elevated blood pressure strains the cardiovascular system and can cause long-term damage if untreated. The Centers for Disease Control and Prevention identifies hypertension as a major contributor to mortality and highlights its strong association with lifestyle and metabolic factors, especially obesity [2], [3]. Clinical studies show that individuals with higher BMI are significantly more likely to develop hypertension, indicating a direct relationship between excess body weight and elevated blood pressure [9].

The association between obesity and hypertension is shaped by numerous interacting factors, including age, gender, cholesterol levels, and additional physiological variables. Conventional clinical methods typically employ fixed thresholds, such as body mass index (BMI) categories or blood pressure cutoffs, to assess risk. Although these thresholds facilitate diagnosis, they do not adequately account for the combined influence of multiple risk indicators. This limitation underscores the necessity for more adaptable approaches capable of simultaneously analyzing several variables to yield a more comprehensive evaluation of health risk [1], [3].

Machine learning provides a data-driven methodology for modeling complex relationships within health datasets. By simultaneously analyzing multiple features, machine learning models estimate the probability of specific outcomes, rather than relying solely on single-variable thresholds. The National Health and Nutrition Examination Survey (NHANES) dataset serves as a reliable source, integrating demographic, clinical, and laboratory measurements [10]. This dataset enables the investigation of how obesity-related features contribute to hypertension risk and other health outcomes, as well as the assessment of the strengths and limitations of predictive modeling across various conditions.

Common machine learning techniques applied in healthcare prediction include Logistic Regression, Neural Networks, and ensemble-based models. Logistic Regression is frequently employed because of its interpretability and effectiveness in binary classification tasks. Neural Networks can model complex, non-linear relationships among variables. Ensemble methods, such as Random Forest, enhance predictive performance by integrating multiple decision trees. The effectiveness of these models is influenced by data quality and the strength of associations between predictors and outcomes, which may differ substantially across various health conditions [11], [12].

Obesity is associated not only with hypertension but also with a range of other health conditions, including kidney disease, stroke, and metabolic disorders. These conditions are interconnected and share common risk factors, such as elevated blood pressure, abnormal cholesterol levels, and impaired glucose metabolism. This broader spectrum of conditions demonstrates that multiple health outcomes can be influenced by shared underlying factors. Understanding these relationships is essential for evaluating the contribution of obesity to hypertension and other health risks within the population [8], [9].

Certain health outcomes, such as hypertension, exhibit strong and direct associations with measurable variables like body mass index (BMI) and blood pressure. In contrast, outcomes such as kidney disease, stroke, and sleep-related disorders involve more complex biological and behavioral factors that are not fully represented in standard clinical datasets. Consequently, predictive models for these conditions may demonstrate lower performance, even when advanced machine learning techniques are applied. This variation in predictive performance underscores the importance of comparing multiple outcomes rather than focusing exclusively on a single condition.

Machine learning offers a robust framework for analyzing complex datasets; however, its effectiveness is highly dependent on the quality and

completeness of the available data. In healthcare applications, factors such as missing variables, measurement variability, and population heterogeneity can significantly affect model performance. Therefore, machine learning models should be evaluated not only for predictive accuracy but also for their limitations and applicability to real-world scenarios. This study employs multiple machine learning approaches using NHANES data to assess the predictive value of obesity-related features for hypertension and other health outcomes, while also identifying the strengths and limitations of these models [11], [12].

Problem Statement

Obesity and hypertension are extensively researched health conditions; however, accurately assessing individual risk remains challenging. Conventional clinical methods utilize fixed thresholds, such as body mass index (BMI) categories and blood pressure cutoffs, to classify individuals as healthy or at risk. Although these thresholds facilitate general screening, they do not adequately represent the complex interactions among multiple health indicators, including age, body composition, and metabolic factors. Consequently, individuals with similar measurements may experience different levels of risk that are not captured by standard classification methods [1], [3], [9].

Beyond these limitations, predictive modeling across multiple health outcomes introduces additional challenges. Some conditions, such as hypertension, are closely associated with measurable variables like BMI and blood pressure, which facilitates more accurate prediction. In contrast, outcomes such as kidney disease, stroke, and sleep-related disorders are influenced by a wider array of biological and behavioral factors that may not be comprehensively represented in available datasets. This discrepancy can result in variability in model performance and diminished predictive accuracy for these outcomes.

A further significant issue is the imbalance and variability inherent in real-world health data. Datasets such as the National Health and Nutrition Examination Survey (NHANES) often contain missing values, uneven class distributions, and demographic differences across population groups, all of which can adversely affect the performance of machine learning models. These challenges underscore the necessity for robust evaluation methods and careful interpretation of results when implementing predictive models in healthcare settings [10], [11].

Accordingly, this study addresses the development and evaluation of machine learning models designed to estimate the probability of obesity-related health risks, with a primary emphasis on hypertension. Furthermore, the study examines how obesity-related features influence variations in predictive performance across multiple health outcomes, thereby providing insights into both the strengths and limitations of machine learning approaches within this domain.

OBJECTIVE

This study aims to develop and evaluate machine learning models to estimate the probability of health risks associated with obesity, with a primary focus on hypertension. Data from the NHANES 2017–2018 dataset are used to analyze how key features, including body mass index, waist circumference, age, and blood pressure, contribute to predicting the risk of hypertension. Additionally, this study compares predictive performance across multiple health outcomes, including kidney risk, stroke, and sleep-related conditions. By evaluating these outcomes with consistent modeling techniques and performance metrics, the study identifies which conditions are more effectively predicted using obesity-related features and examines the strengths and limitations of machine learning approaches in health risk prediction.

METHODOLOGY

This study utilizes data from the National Health and Nutrition Examination Survey (NHANES) 2017–2018 dataset, which provides a

comprehensive set of demographic, clinical, and laboratory measurements representative of the U.S. population [10]. The dataset was constructed by merging multiple data files using a common participant identifier, resulting in a unified dataset containing variables relevant to obesity and related health outcomes.

Key features selected for analysis included body mass index (BMI), waist circumference, age, gender, and blood pressure measurements. Additional variables, including averages of systolic and diastolic blood pressure, cholesterol levels, and glucose measurements, were incorporated to improve predictive performance. Several binary outcome variables were defined, such as hypertension, kidney risk, stroke, and sleep-related conditions, based on clinical thresholds and questionnaire responses.

Data preprocessing followed a structured pipeline approach. Missing values in numerical features were addressed using median imputation, while categorical variables underwent most-frequent imputation and subsequent one-hot encoding. Numerical features were standardized through feature scaling to ensure consistent input ranges across models.

Several machine learning models were implemented to assess predictive performance. Logistic Regression served as the baseline model because of its interpretability and effectiveness in binary classification tasks [11]. Additionally, a Neural Network model was employed to capture potential nonlinear relationships among features, while a Random Forest classifier was utilized to investigate ensemble-based modeling techniques. These models were chosen to facilitate comparison between linear, non-linear, and ensemble approaches [11], [12].

Model evaluation employed 5-fold stratified cross-validation to enhance robustness and address class imbalance in the dataset. The primary performance metric was the area under the receiver operating characteristic curve (ROC-AUC), which quantifies the model's ability to distinguish between positive and negative cases [11]. Additional metrics, such as precision, recall, and F1-score, were calculated to provide a more comprehensive assessment of classification performance.

Threshold tuning was applied to the predicted probabilities to enhance classification results. Rather than relying exclusively on the default threshold of 0.5, multiple threshold values were evaluated to determine the optimal balance between precision and recall, particularly for imbalanced outcomes. This strategy resulted in improved F1 Scores when prioritizing the detection of positive cases.

Overall, this methodology establishes a systematic framework for evaluating the contribution of obesity-related features to predicting multiple health outcomes. It also enables comparison of various machine learning approaches using consistent evaluation criteria.

MODEL LIMITATIONS

The findings indicate that machine learning models effectively predict hypertension risk using obesity-related features but demonstrate lower predictive accuracy for other health outcomes (see Figure 1).

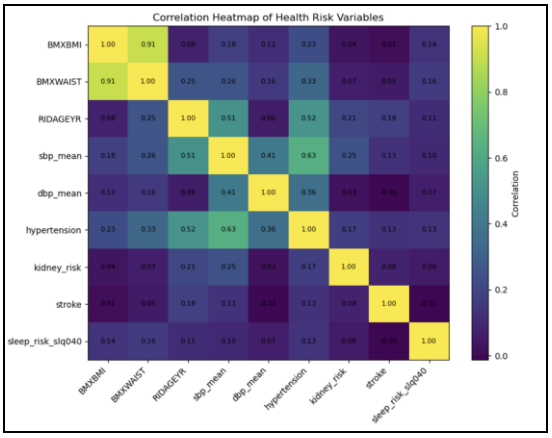

Figure 1
Correlation of Health Risk Variables

Hypertension is closely linked to measurable variables such as body mass index (BMI) and blood pressure, which contribute to higher predictive accuracy. In contrast, conditions including kidney disease, stroke, and sleep-related disorders are

influenced by additional biological and behavioral factors not fully represented in the dataset. This limitation reduces the models' ability to accurately predict these outcomes.

Furthermore, challenges such as class imbalance, missing data, and variability in health measurements may have further affected model performance. These limitations underscore the necessity for cautious interpretation of machine learning results in healthcare contexts [10], [11].

Despite these challenges, the study demonstrates that machine learning offers valuable insights into obesity-related health risks and identifies areas where additional data and enhanced modeling approaches are required.

Relationship Between Obesity and Hypertension

The relationship between obesity and hypertension was analyzed using both graphical and statistical approaches. Figure 2 illustrates the relationship between body mass index (BMI) and systolic blood pressure, showing a general upward trend as BMI increases. The results demonstrate that individuals with higher BMI values tend to exhibit increased systolic blood pressure, indicating a positive association between obesity and hypertension risk.

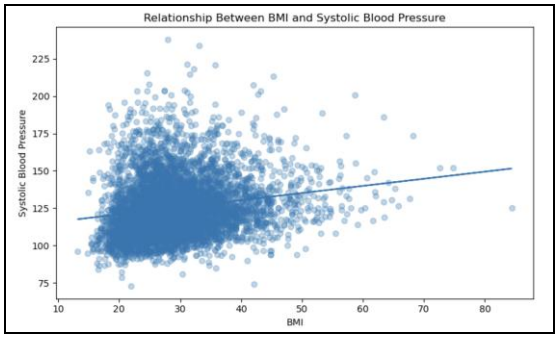

Figure 2
Relationship Between BMI and Systolic Blood Pressure

Figure 3 provides additional evidence by comparing hypertension prevalence between obese and non-obese groups. The analysis reveals that individuals classified as obese exhibit a higher proportion of hypertension cases than non-obese individuals. This result aligns with previous clinical research that identifies obesity as a significant risk factor for elevated blood pressure [9].

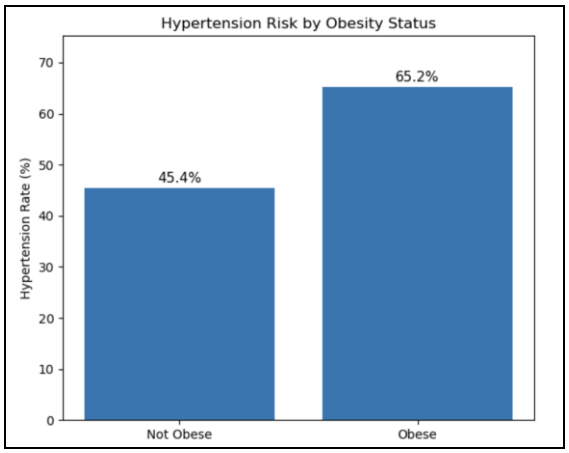

Figure 3
Comparison Between the Obese and Non-Obese Groups

These findings confirm that obesity-related characteristics, particularly BMI and waist circumference, are strongly associated with hypertension risk. This association accounts for the superior predictive performance of machine learning models for hypertension relative to other outcomes.

Analysis of Results

The performance of machine learning models varied across health outcomes, with hypertension exhibiting the highest predictive accuracy among all targets. Figure 4 demonstrates that the logistic regression model achieved the strongest performance for hypertension prediction, with a ROC-AUC of approximately 0.80-0.82, indicating a strong ability to distinguish positive from negative cases. In contrast, models for stroke and kidney disease achieved moderate performance, with ROC-AUC values ranging from approximately 0.72 to 0.78. Sleep-related risk prediction yielded substantially lower performance, with ROC-AUC values near 0.58 to 0.60.

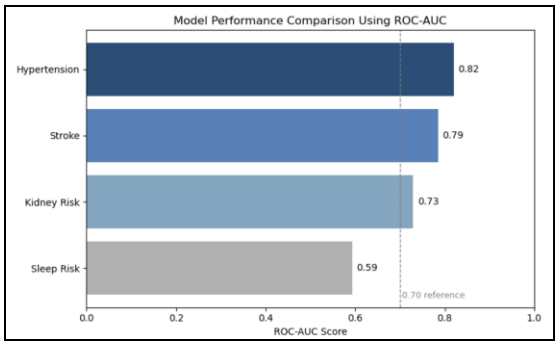

Figure 4
Model Performance Comparison Using ROC-AUC

The superior performance in hypertension prediction is attributable to the direct availability of highly relevant clinical features, particularly systolic and diastolic blood pressure measurements, as well as obesity-related indicators such as body mass index (BMI) and waist circumference. These variables provide strong, direct signals that are closely aligned with the underlying condition, enabling the model to identify patterns more effectively. Conversely, outcomes such as stroke, kidney disease, and sleep disorders are influenced by more complex and multifactorial processes that are not fully captured by the available dataset features.

Across all outcomes, logistic regression demonstrated performance comparable to, and in some cases more stable than, more complex models such as neural networks and random forest classifiers. Although advanced models can capture non-linear relationships, their performance gains in this study were limited, likely due to the relatively small feature set and the data's structured nature. Consequently, logistic regression served as a reliable and interpretable baseline model for predicting health risks in this context.

CONCLUSION

This study showed that machine learning techniques can effectively model health risk probabilities using NHANES 2017–2018 data. Hypertension had the highest predictive performance, likely because it is linked to clear clinical features like blood pressure and obesity-related indicators. On the other hand, predictions for stroke, kidney disease, and sleep disorders were less accurate, which may be due to the complex and varied causes of these conditions.

The analysis also confirmed a strong link between obesity and hypertension. Both the graphs and statistics showed that people with higher body mass index (BMI) and waist circumference are more likely to have high blood pressure and a greater risk of hypertension. These results match previous clinical research and highlight obesity as an important risk factor for heart health.

In summary, the results demonstrate that although machine learning models can effectively capture patterns in structured health data, their performance is highly contingent upon the availability of relevant and high-quality features. Logistic regression emerged as a reliable and interpretable model in this context, performing comparably to more complex approaches. These findings emphasize the potential of data-driven methods to support early risk detection and highlight the necessity for comprehensive datasets to enhance predictive accuracy across diverse health outcomes.

FUTURE WORK

Future research should aim to improve model performance by incorporating additional variables that capture behavioral, environmental, and longitudinal health factors currently absent from the dataset. Including lifestyle-related data such as diet, physical activity, sleep quality, and medical history may substantially enhance the predictive capability of models, especially for complex conditions such as stroke, kidney disease, and sleep disorders. In addition, exploring advanced machine learning techniques, such as ensemble methods and deep learning architectures, may facilitate the identification of non-linear relationships that traditional models do not fully capture. Further research should also consider hyperparameter optimization and feature selection techniques to

refine model performance and minimize the impact of data noise.

Another important direction is the use of longitudinal datasets to improve understanding of how health risks change over time. Because NHANES data is cross-sectional, it restricts the ability to establish temporal relationships between variables. Incorporating time-series data would enable more accurate modeling of disease progression and risk prediction.

Finally, future research should address model interpretability and real-world application, including the development of decision-support tools for healthcare professionals. Ensuring that models are both accurate and explainable is essential for their adoption in clinical environments and for supporting informed healthcare decisions.

REFERENCES

[1] Centers for Disease Control and Prevention. (2024, May 14). *Adult Obesity Facts* [Online]. Available: https://www.cdc.gov/obesity/adult-obesity-facts/index.html.

[2] World Health Organization. (2025, September 25). *Hypertension* [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/hypertension.

[3] National Center for Health Statistics. (2024, August 7). *Hypertension—Health, United States: Sources and Definitions* [Online]. Available: https://www.cdc.gov/nchs/hus/topics/hypertension.htm.

[4] National Center for Health Statistics, "2017–2018 Data Documentation, Codebook, and Frequencies: Blood Pressure (BPX_J)," *Centers for Disease Control and Prevention*, Hyattsville, MD, Feb. 2020.

[5] National Center for Health Statistics, "2017–2018 Data Documentation, Codebook, and Frequencies: Medical Conditions (MCQ_J)," *Centers for Disease Control and Prevention*, Hyattsville, MD, Feb. 2020.

[6] National Center for Health Statistics, "2017–2018 Data Documentation, Codebook, and Frequencies: Sleep Disorders (SLQ_J)," *Centers for Disease Control and Prevention*, Hyattsville, MD, Feb. 2020.

[7] World Health Organization. (2024). *Obesity and Overweight* [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight.

[8] Centers for Disease Control and Prevention. (2023). *Adult Obesity Prevalence Maps* [Online]. Available: https://www.cdc.gov/obesity/data-and-statistics/adult-obesity-prevalence-maps.html.

[9] J. Hall et al., "Obesity-induced hypertension: interaction of neurohumoral and renal mechanisms." *Circulation Research*, vol. 116, no. 6, pp. 991–1006, Mar. 2015.

[10] National Center for Health Statistics, "National Health and Nutrition Examination Survey: Overview 2017–2018," Centers for Disease Control and Prevention, Hyattsville, MD, 2020.

[11] T. Hastie et al., "Supervised learning," in *The Elements of Statistical Learning*, 2nd ed. New York, NY, USA: Springer, 2009, ch. 2, pp. 9–41.

[12] I. Goodfellow et al., *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.