

## Leveraging Machine Learning for Predicting Health Perceptions: A Behavioral Analytics Approach Using U.S. National Survey Data

Rezwan Ahmed<sup>1,\*</sup>  
Arnob Zahid<sup>2</sup>

\*Corresponding author

<sup>1</sup> Graduate Student, BRAC University, Dhaka, Bangladesh, rezwan.ahmed1@g.bracu.ac.bd

<sup>2</sup> Lecturer, University of Waikato, Hamilton, New Zealand, arnob.zahid@waikato.ac.nz

### ABSTRACT

This study leverages machine learning to predict self-rated health (SRH) among U.S. adults using the 2023 BRFSS dataset (n = 117,386). SRH was framed as a binary outcome (Good vs. Poor) and modeled using Logistic Regression, Random Forest, XGBoost, Decision Tree, and a stacked ensemble. To address class imbalance, ENN, SMOTE, and repeated under-sampling were compared, with repeated under-sampling yielding the most consistent performance. Logistic Regression achieved an AUC of 0.81, with the stacked model slightly improving to 0.82. SHAP analysis revealed mental health, exercise, income, and diabetes as top predictors. Findings, interpreted through the Social Determinants of Health and Health Belief Model, emphasize socioeconomic and behavioral influences on perceived health.

**Keywords:** Self-rated health, machine learning, imbalanced data, health analytics, behavioral risk factor surveillance system (BRFSS).

### INTRODUCTION

Self-rated health (SRH) is a widely used indicator in the public health research discipline and often employed to assess population health and predict long term health outcomes. It is a simple and easy to use measurement for assessing people's general health. It is usually applied by asking "Would you say that in general your health is?", and in response, to rate their health on a scale like "excellent", "very good", "good", "fair" or "poor" (Mu *et al.*, 2024; Wuorela *et al.*, 2020). SRH is a comprehensive and dynamic assessment that encompasses more factors than typical covariant studies. It influences health-related behaviors and reflects an individual's ability to cope with future health challenges (Idler & Benyamini, 1997; Subramanian *et al.*, 2010). It can capture a person's overall perception of their well-being (World Health Organization, 2004). SRH is important because perception of health can be a strong predictor of future health outcomes, lifestyle choices, and even mortality rates, despite its non-specific nature (Lorem *et al.*, 2020; Wuorela *et al.*, 2020). Understanding factors that influence an individual's health perceptions can be crucial for public health interventions and policies because SRH captures both subjective and contextual dimensions.

In recent years, growing large-scale survey datasets such as the Behavioral Risk Factor Surveillance System (BRFSS) of U.S. and advancement of machine learning (ML) have created new opportunity to model SRH and identify its key determinants. Machine learning is enabling the detection of complex relationships between demographic, behavioral, and socioeconomic variables in large datasets that traditional statistical models might overlook. However, several challenges remain. First, predicting SRH from national survey data involves methodological issues such as class imbalance, which can bias models toward majority responses. Second, while ML is increasingly applied in health research, many studies emphasize predictive performance at the expense of interpretability and rarely connect findings to established theoretical frameworks. This limits their utility for guiding interventions. Third, transparency in reporting particularly regarding model performance across outcome classes and reproducibility remains inconsistent.

To address these gaps, this study focuses on three research questions: (1) Which socio-demographic, behavioral, and clinical variables are the most important predictors of self-reported health among U.S. adults in the 2023 BRFSS? (2) How do different machine learning models compare in predicting SRH? (3) Can the findings be linked to established frameworks such as the Social Determinants of Health (SDOH) and the Health Belief Model (HBM) to strengthen theoretical and practical relevance? By applying ML methods to recent BRFSS data and interpreting results with public health theory, this study seeks to provide reproducible, evidence-based insights into the factors shaping perceived health in the U.S. population.

### LITERATURE REVIEW

The literature has long established self-rated health as a robust and multidimensional indicator of health outcomes. It is a subjective evaluation of an individual and an understanding of their disease burden. It is also an expectation for their overall health status. Extensive research indicates that SRH has a strong predictive power for an individual's risk of death and there is significant variety in the responses to SRH among different populations (Mu *et al.*, 2024). The determinants of SRH include

societal characteristics and individual socioeconomic, biomedical, and behavioral factors. It is also influenced by individual subjective conceptualizations of their built environment context, demonstrating an interplay between social and environmental determinants (Parga *et al.*, 2024). Numerous studies have examined the factors influencing SRH using survey data. A broad set of factors influence SRH. Socioeconomic indicators such as income, education, and employment are consistently associated with better perceived health as financial stability and access to resources are often associated with better health outcomes (Dierx & Kasper, 2022; Weech-Maldonado *et al.*, 2017). Behavioral factors such as physical activity, diet, smoking, and alcohol consumption play a strong role in shaping health perceptions (Barreto & Figueiredo, 2009). Clinical conditions, such as hypertension, diabetes, and cancer history, also affect SRH, though their impact can vary by demographic subgroup (Latham & Peek, 2013). Psychosocial variables, particularly mental health status and depression play a crucial role, as individuals experiencing psychological distress tend to rate their health lower (Ozcakir *et al.*, 2014; Tadiri *et al.*, 2021).

However, there remains an ongoing question about the relative importance of these factors (Balaj, 2022; Gallagher *et al.*, 2016; Hamplová *et al.*, 2022; Linton *et al.*, 2016). Traditional statistical methods may not fully capture the complex interactions or nonlinearities in large health datasets (Chowdhury *et al.*, 2023; Desai *et al.*, 2020; Liu *et al.*, 2021). So, machine learning has been increasingly used to predict SRH and health related outcomes which is offering greater flexibility in capturing complex relationships among predictors (e.g., Choi & Jung, 2025; El Morr *et al.*, 2024; Gumà-Lao & Arpino, 2023; Hoekstra *et al.*, 2023; Ku & Min, 2024), but many of these studies either lack model interpretability or do not connect empirical results to broader theoretical perspectives. Moreover, health survey data often exhibit imbalanced distributions of SRH categories, with “Good/Excellent” responses far outnumbering “Fair/Poor.” (Chowdhury *et al.*, 2024) This class imbalance can bias models toward majority-class predictions, reducing sensitivity to poorer health outcomes. So, evaluating sampling methods with class-specific metrics is crucial for fair model assessment.

To contextualize the findings, the study draws upon two well-established public health theories: the Social Determinants of Health (SDOH) framework and the Health Belief Model (HBM). The SDOH is endorsed by the World Health Organization (WHO) and it emphasizes that the health outcomes and perceptions are shaped by non-medical factors such as income, education, employment, social support, and living conditions and environment (Chaturvedi *et al.*, 2024; Organization, 2024). Addressing the SDOH is crucial for improving health outcomes and reducing health disparities. Complementing SDOH with Health Belief Model (HBM) which offers a psychological lens to understand how individuals evaluate their health. HBM is a psychological framework which explains health related behaviors. It suggests that individuals' health behaviors and perceptions are influenced by their beliefs about their susceptibility to a disease, the severity of the disease, the benefits of taking action, and the barriers to taking action (Alyafei & Easton-Carr, 2025; Janz & Becker, 1984; Limbu *et al.*, 2022; Suess *et al.*, 2022). Mapping ML-derived predictors to these frameworks bridges data-driven insights with established theory, enhancing policy relevance.

## MATERIALS AND METHODS

This study adopts the following methodology to explore the determinants of SRH using interpretable machine learning. The methodology is illustrated in Figure 1.

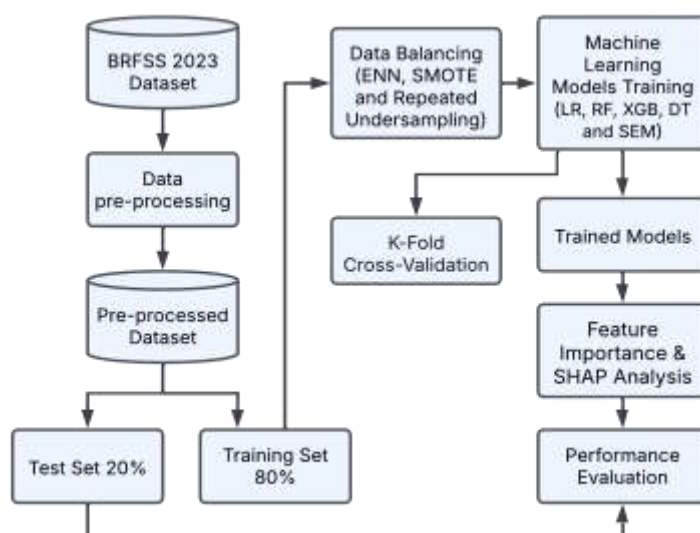


Figure 1: Overview of the study process from data collection to interpretation.

The BRFSS dataset of 2023 was used in this study. Main processes include: data preprocessing, data balancing, model training, and SHAP analysis and performance evaluation. Initially, the dataset was cleaned and prepared for analysis. Furthermore, to

address the class imbalance of the target variable, data balancing techniques such as ENN, SMOTE and Repeated under-sampling are applied to identify the most effective technique. Machine learning models including, logistic regression, random forest, XGBoost, decision trees and stacked ensemble model were trained and validated using k-fold cross-validation to ensure robust model evaluation and reduce overfitting. Finally, model performance was examined using evaluation metrics to determine the effective approach. All the analyses and ML model training were conducted using Python programming language on a cloud-based platform, Google Colab (Google, n.d.). Google Colab provided free GPU support and ease of collaboration for the computational experiments.

### Data Source

The BRFSS survey is a nationally recognized annual survey that collects data on the health status of American residents. The survey is run by the Centers for Disease Control and Prevention (CDC) and is conducted via telephone. The BRFSS provides a rich and nationally representative dataset of U.S. adults and it contains a wide range of factors related to health behavior, demographics, healthcare access and chronic conditions of a diverse population thus providing a valuable opportunity to explore the multifaceted determinants of SRH (Chowdhury *et al.*, 2024; Imoisili *et al.*, 2024; Sekkarie *et al.*, 2024). It has been instrumental in many studies because of its reliability and comprehensive data (Nelson *et al.*, 2001; Pierannunzi *et al.*, 2013). The BRFSS of 2023 survey consists of records of 433,324 American residents featuring 350 attributes (Centers for Disease Control and Prevention [CDC], 2023). For our SRH classification, the ‘General Health’ feature was selected as the target variable. Participants of the BRFSS were asked to report their general health in this feature (Table 1). This SRH question is formulated as “Would you say that in general your health is:” Participants rated their general health as ‘Excellent’, ‘Very good’, ‘Good’, ‘Fair’ or ‘Poor’. The answer captures people's own assessment of health according to their own definition of health. It is a subjective yet powerful predictor of mortality in the general population (Idler & Benyamini, 1997; Latham & Peek, 2013). We excluded participants who “did not respond” or “refused to answer” from the study to ensure data completeness and minimize non-response bias. The study considered a comprehensive set of features surrounding demographic, behavioral, psycho-social, chronic conditions, and healthcare-related factors. The inclusion of these categories was guided by existing literature. Based on insights drawn from a comprehensive review of prior research (as referenced in literature review), 19 independent features were selected with self-reported health as the dependent outcome. A detailed list of the selected features is presented in Table 2 with their values and counts after preprocessing.

Table 1: General health question and response options in the BRFSS.

Question	Response	Value
“Would you say that, in general, your health is...?”	Excellent	1
	Very good	2
	Good	3
	Fair	4
	Poor	5
	Don't know/Not Sure	7
	Refused	9

Note: In Table 1, The response values are numerically coded in the BRFSS dataset, as shown in the 'Value' column.

Table 2: Summary statistics of the dataset after preprocessing.

Feature	Value	Count	Percentage (%)
General health (Target variable)	Good	89,279	76.1
	Poor	28,107	23.9
Age group	18-24	1,970	1.7
	25-34	9,080	7.7
	35-44	16,273	13.9
	45-54	17,167	14.6
	55-64	23,067	19.6
	65+	49,829	42.5
Sex	Male	63,294	53.9
	Female	54,092	46.1
Race/Ethnicity	White, non-Hispanic	94,296	80.3
	Black, non-Hispanic	6,586	5.6
	Asian, non-Hispanic	1,700	1.5
	American Indian, Non-Hispanic	2,377	2.1
	Hispanic	7,747	6.6
	Other race, non-Hispanic	4,680	3.9
Education level	Grade 1 through 11	7,249	6.2
	Grade 12 or GED (High school graduate)	33,532	28.6
	College 1-3 years	36,703	31.3
	College 4+ years (College graduate)	39,902	34.0

Feature	Value	Count	Percentage (%)
Employment	Employed	55,971	47.7
	Unemployed/Retired/homemaker/student	61,415	52.3
Annual income	< 15k	7,678	6.5
	15-25k	12,286	10.5
	25-35k	14,476	12.3
	35-50k	17,463	14.9
	50-100k	36,454	31.1
	100-200k	22,663	19.3
Exercise	> 200k	6,366	5.4
	Yes	84,742	72.2
Smoke	No	32,644	27.8
	Every day/Some days	33,148	28.3
Heavy drinkers	Not at all	84,238	71.7
	Yes	11,458	9.8
Diabetes	No	105,928	90.2
	Yes	19,002	16.2
	Pre-diabetes	3,145	2.7
Heart attack	Yes	9,645	8.2
	No	107,741	91.8
Stroke	Yes	6,649	5.7
	No	110,737	94.3
Melanoma	Yes	16,725	14.2
	No	100,661	85.8
BMI	Underweight	1,956	1.7
	Normal weight	31,811	27.1
	Overweight	42,550	36.2
	Obese	41,069	35.0
Depression	Yes	30,429	25.9
	No	86,957	74.1
Mental health	1-30	49,904	42.5
	0	67,482	57.5
Healthcare provider	Yes	104,291	88.8
	No	13,095	11.2
Checkup	Regular	95,874	81.7
	Irregular	21,512	18.3

Note: In Table 2, General health is the self-reported health status and serves as the target variable in this study. Melanoma refers to Melanoma or any other type of cancer. Mental health represents the number of days in the past month when an individual's mental health was not good, ranging from 1 to 30 days.

### Data Pre-processing and Class-Balancing

In the stage of pre-processing, records with missing values were removed to ensure data integrity. Additionally, instances where participants responded 'Don't Know/Not Sure' and 'Refused' were also removed. No data synthesis or data imputation were applied as the study aimed to train the models using the actual collected data. All selected BRFSS variables were in integer-coded categories in the raw dataset. Variables with a natural order (e.g., education level, income, poor mental health) were retained in their numeric form to preserve ordinal information. For nominal variables without a natural order (e.g., race/ethnicity, marital status, sex), we performed one-hot encoding. After conducting pre-processing, a total of 117,386 records were retained for analysis and model training. For our binary classification approach, we grouped 'Excellent', 'Very good' and 'Good' under 'good health' and 'Fair' and 'Poor' under 'poor health'.

As part of the preprocessing steps, we conducted a descriptive analysis of the dataset to assess the distribution of the features (Table 2). In terms of demographics, the majority of the participants fall in the 65+ age group (42.5%) and only few (1.7%) are in the 18-24 age group. Regarding gender, the dataset was fairly balanced with 53.9% males and 46.1% females. The participants were mostly white and non-Hispanic (80.3%). Regarding education, 34% of the participants have completed college whereas 6.2% have less than a high school education. 47.7% participants are employed and 52.3% either unemployed, retired or students. 55.8% of participants belong to higher-income groups, earning USD 50K or more. In terms of health behaviors, 72.2% reported engaging in physical exercise. 28.3% reported smoking daily or occasionally and 9.8% are heavy alcohol consumers. 25.9% reported having depression which shows a notable portion of the sample are with mental health concerns. Among chronic conditions, 16.2% have diabetes, 8.2% have had a heart attack and 5.7% have had a stroke. In terms of healthcare access, 81.7%

of the participants reported doing regular checkups, while 18.3% are irregular. Also, 88.8% of the participants have a personal healthcare provider which suggests most respondents have access to healthcare.

During the descriptive analysis, significant class imbalance was found in the target variable 'General health' (Table 2). 76.06% participants reported being in good health, whereas only 23.94% reported being in poor health. To address this class imbalance, both over-sampling and under-sampling techniques were applied. Methods including, Edited Nearest Neighbors (ENN), Synthetic Minority Over-sampling Technique (SMOTE), and Repeated Under-sampling were employed on the training dataset to ensure a more balanced representation of the target classes and optimum models' performance. ENN removes noisy or ambiguous samples by eliminating instances from the training dataset which are potential to increase the model's accuracy (Alejo *et al.*, 2010; Halder *et al.*, 2024). On the other hand, SMOTE generates synthetic data of the minority class to balance the unbalanced dataset. Synthetic data points are created by interpolating between the original instances and the neighbors. SMOTE is effective in improving classification performance on imbalanced datasets (Chawla *et al.*, 2002; Joloudari *et al.*, 2023; Wongvorachan *et al.*, 2023). All Sampling techniques were done only on the training folds during cross-validation to avoid leakage.

### ***Repeated under-sampling***

A Repeated Under-sampling approach was implemented to address the class imbalance (Haibo He & Garcia, 2009; Xu-Ying Liu *et al.*, 2009). In traditional under-sampling techniques instances are randomly removed from the majority class, which can lead to loss of valuable information. Therefore, to mitigate this, the dataset was first divided into two subsets based on the target variable (Poor health and good health). Then, the majority class was randomly under-sampled five times. Five distinct balanced datasets were created and each contained the full minority class and a randomly selected subset of the majority class. For each balanced dataset, we separated the data into training and test sets using stratified sampling and trained a model on the training data. Furthermore, performance metrics from all iterations were aggregated to obtain a more robust and reliable evaluation. This method enhances model stability by reducing the variance introduced by a single random under-sampling and provides a more generalized performance estimate.

## **Machine Learning Model Development and Evaluation**

### ***Model training***

Logistic Regression (LR), Random Forest (RF), Extreme Gradient Boosting (XGB) and Decision Trees (DT) models were chosen to train on the dataset. These classification models have proven their performance in similar tasks in previous studies (Chen *et al.*, 2024; Fregoso-Aparicio *et al.*, 2021; Gündoğdu, 2023; Olisah *et al.*, 2022; Sabouri *et al.*, 2023; Zuo *et al.*, 2023). Default settings were used for all models. All ML analyses were performed by Python 3.11. The dataset was split into training and testing sets in an 8:2 ratio, ensuring a balanced evaluation. Also, the training was carried out using 5-fold cross-validation to ensure the model's robustness and generalization. Furthermore, a Stacked Model was implemented. Stacking is an ensemble learning technique where it leverages the strengths of multiple base classifiers and combines them (Ghasemieh *et al.*, 2023; Lu *et al.*, 2023). The stacking ensemble learning consisted of two phases: base models training phase and meta-model training phase. In the base models training phase, LR, RF, XGB, and DT were used as the base models. Each model was trained separately using the same training data. In the meta-model training phase, the predictors generated by the base models were used as inputs to train a Logistic Regression model, which acted as the meta-model.

### ***SHAP analysis***

To establish a deeper understanding of the predictive model, Shapley Additive exPlanation (SHAP) values were calculated from the best performing model which assisted in determining how each feature influenced the model's predictions. A summary plot was created to present how each feature has contributed in the output. SHAP was introduced by Lundberg and Lee and has gained popularity as a method for interpreting machine learning model predictions. SHAP provides insights into the contribution of each feature to specific predictions by utilizing Game Theory techniques (Lundberg & Lee, 2017; Wang *et al.*, 2024). SHAP values are calculated to decompose predictions into additive contributions from each feature. This approach provides both global interpretability meaning identifying the overall influential features and local explanations by revealing how individual features shift predictions for specific instances. The method rigorously attributes model output changes to each variable while controlling for feature interactions. It ensures a mathematically consistent assessment of each predictor's role in the model's decision logic. In this study, for the tree-based models, SHAP Tree Explainer of SHAP library in Python were utilized. And for logistic regression, SHAP Linear Explainer was utilized which works for linear relationships. It explained the coefficients in terms of contributions.

### ***Performance evaluation & comparison***

The trained models were evaluated using several standard classification metrics. Accuracy (Equation 1), precision (Equation 2), recall (Equation 3), F1-score (Equation 4), and Area Under the Receiver Operating Characteristic Curve (AUC-ROC) were estimated (Equation 5). Additionally, the mean cross-validation (CV) AUC was also computed. Accuracy is the proportion of correctly classified instances over the total number of samples, while precision measures the percentage of predicted positive cases that are actually positive. Recall (or sensitivity) indicates the percentage of actual positives correctly identified. F1-score is the harmonic mean of precision and recall. The AUC-ROC shows the model's ability to distinguish between positive and negative classes. It is computed from the ROC curve which plots the true positive rate against the false positive rate at various thresholds (Google Developers, n.d.). The models were evaluated after being trained using stratified 5-fold cross-validation on

the training data then the mean CV AUC was calculated to see consistency across folds and to prevent any overfitting. These metrics provided a comprehensive view of model effectiveness. The best-performing model was selected primarily based on the largest AUC while also considering the balance of other metrics. All metrics were computed using built-in functions from the scikit-learn library in Python programming language. The evaluation metrics are defined as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$F1 - score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad (4)$$

$$ROC - AUC = \int_0^1 TPR(FPR) dFPR \quad (5)$$

Here, TP: True Positive, TN: True Negative, FP: False Positive, FN: False Negative, TPR: True Positive Rate, FPR: False Positive Rate.

## RESULTS

### Model Performance

The BRFSS 2023 dataset used in this study was heavily imbalanced for model training. To address this data imbalance, both over-sampling and under-sampling techniques were tested on the training data (Figure 2 and Table 3). ENN reduced the number of samples in the majority class (participants with self-rated good health), SMOTE over-sampled the minority class (participants with self-rated poor health) and Repeated Under-sampling used five distinct balanced datasets that were created and each contained the full minority class and a randomly selected subset of the majority class. An LR model was trained to evaluate the performance of the sampling techniques. With the original imbalanced dataset, the F1-score for class 0 (participants with self-rated poor health) is only 47% while for class 1 it's 88%. After applying different sampling techniques, the F1-score for class 0 improves to 57% with ENN, 55% with SMOTE and 73% with Repeated Under-sampling. Finally, the dataset processed with Repeated under-sampling was selected as it showed the most balanced performance across both classes. After addressing the class imbalance, ML models were trained and their performance was evaluated. Figure 3 and Table 4 summarize the performance metrics for Logistic Regression, Random Forest, XGBoost, Decision Tree, and the Stacked Model. Among the models, the Stacked Model and Logistic Regression demonstrated the strongest performance. LR achieved AUC of 81% and F1-score of 73%. Stacked model attained a slightly higher AUC of 81.8% and F1-score of 74%. In contrast, RF, XGB and DT exhibited lower performance metrics. These results were consistent across cross-validation folds and demonstrated a balanced performance in predicting perceived general health.

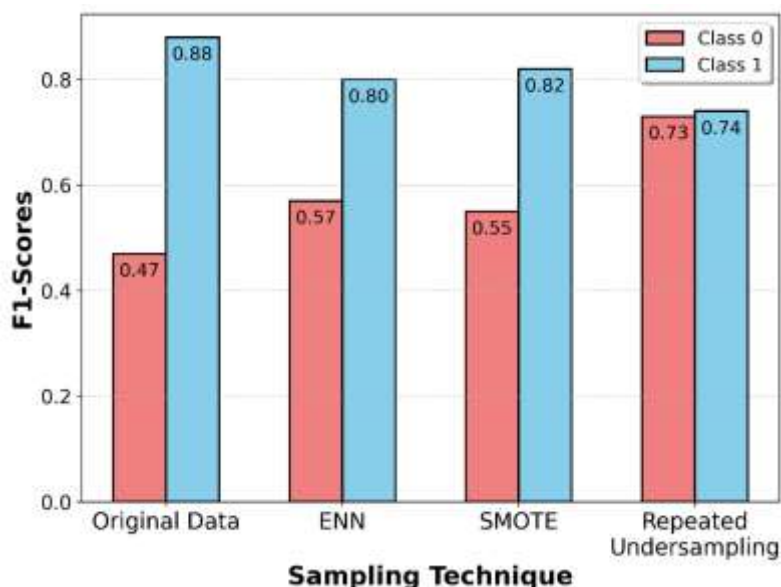


Figure 2: Comparison of F1-Scores Across Sampling Techniques.

Table 3: Performance Comparison of Sampling Techniques Using Logistic Regression.

Sampling Technique	Class	Precision	Recall	F1-Score
Original Data	0	0.66	0.37	0.47
	1	0.83	0.94	0.88
ENN	0	0.50	0.66	0.57
	1	0.88	0.79	0.83
SMOTE	0	0.47	0.67	0.55
	1	0.88	0.77	0.82
<b>Repeated Under-sampling</b>	<b>0</b>	<b>0.74</b>	<b>0.71</b>	<b>0.73</b>
	<b>1</b>	<b>0.72</b>	<b>0.75</b>	<b>0.74</b>

Note: In Table 3, Class 0 represents poor health, while Class 1 represents good health. Bolded values indicate significant metrics. ENN: Edited nearest neighbor; SMOTE: Synthetic minority over-sampling technique.

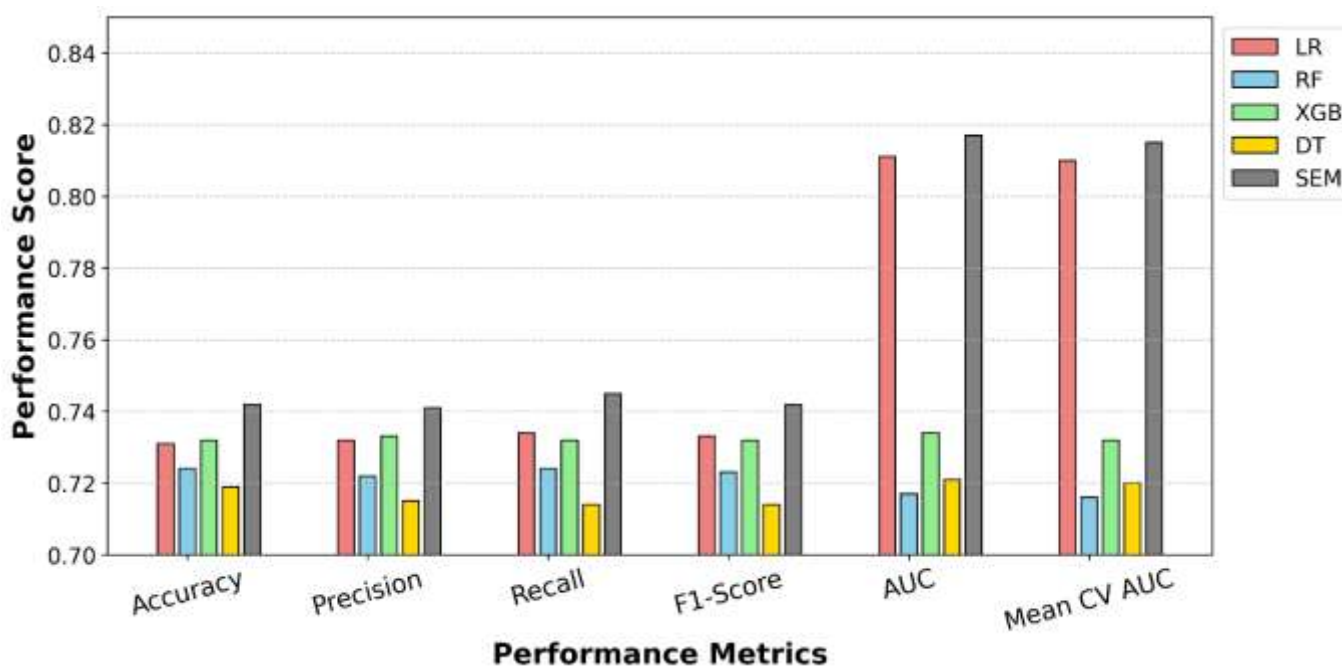


Figure 3: Comparative performance evaluation of multiple predictive models using key metrics. LR: Logistic Regression, RF: Random Forest, XGB: eXtreme Gradient Boosting, DT: Decision Tree, SEM: Stacked Ensemble Model.

Table 4: Performance metrics for various machine learning models trained on BRFSS 2023 Dataset.

Model	AUC	Mean CV AUC	Accuracy	Precision	Recall	F1 Score
<b>Logistic Regression</b>	<b>0.811</b>	<b>0.810</b>	<b>0.731</b>	<b>0.732</b>	<b>0.734</b>	<b>0.733</b>
Random Forest	0.717	0.716	0.724	0.722	0.724	0.723
XGBoost	0.734	0.732	0.732	0.733	0.732	0.732
Decision Tree	0.721	0.720	0.719	0.715	0.714	0.714
<b>Stacked Model</b>	<b>0.818</b>	<b>0.815</b>	<b>0.742</b>	<b>0.741</b>	<b>0.745</b>	<b>0.742</b>

Note: In Table 4, Bold entries are significant (metrics using Logistic Regression and Stacked Model). AUC represents the area under the ROC curve. Mean CV AUC reflects the average AUC obtained across 5-fold cross-validation.

### Model Interpretation With SHAP

Shapley Additive Explanation (SHAP) method was employed to understand models' results. The SHAP analysis provided clear insights into the model's decision logic through two complementary visualizations given (Figure 4 and Figure 5). The SHAP Linear Explainer of Python were used to calculate the SHAP values of logistic regression (LR) as LR performed the best among all singular models. Figure 4 reveals the directional impact of features on predictions. In the plot, the larger the absolute value of a feature's SHAP, the greater hint that this feature has greater impact. Red dots represent the higher values for the feature,

while blue dots represent lower values for the feature. Poor mental health exhibits a pronounced negative influence, indicating its strong association with the outcome. Exercise Frequency with high feature value demonstrates positive SHAP value and with low feature value, it demonstrates negative SHAP value. Similarly, Income Level, Employment, and Diabetes shows profound effect on the output. To complement Figure 4’s plotting, Figure 1 quantifies the overall feature importance by calculating the mean absolute SHAP. Bad Mental Health, Exercise Frequency, Income Level, Employment, Diabetes are the top 5 predictors. On the other hand, Health Care Provider, Heavy Drinking, Routine Checkup, Stroke, and Smoking Frequency had minimal influence, indicating that these factors may be crucial for clinical health outcomes but their influence on self-reported health perceptions may be more subtle.

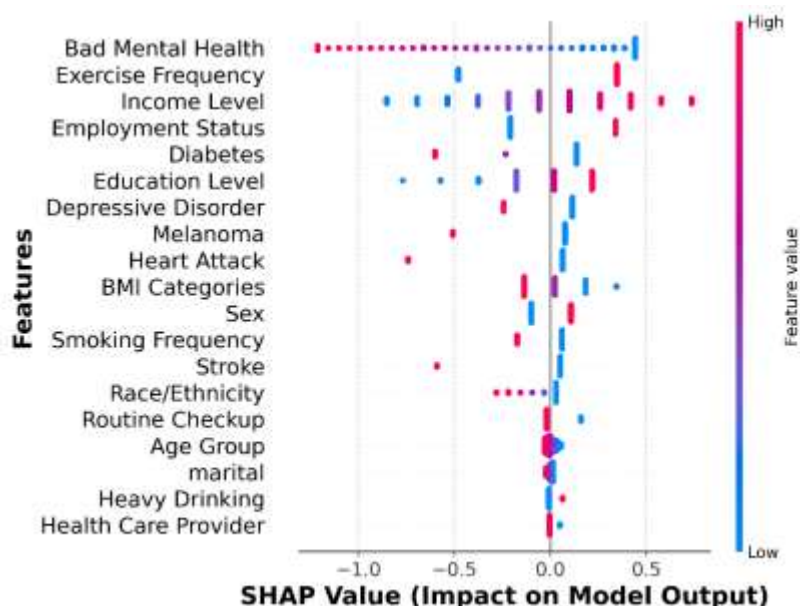


Figure 4: SHAP (Shapley Additive Explanation) values demonstrating the impact of individual features on model predictions. The colors represent the feature values: red for larger values and blue for smaller values.

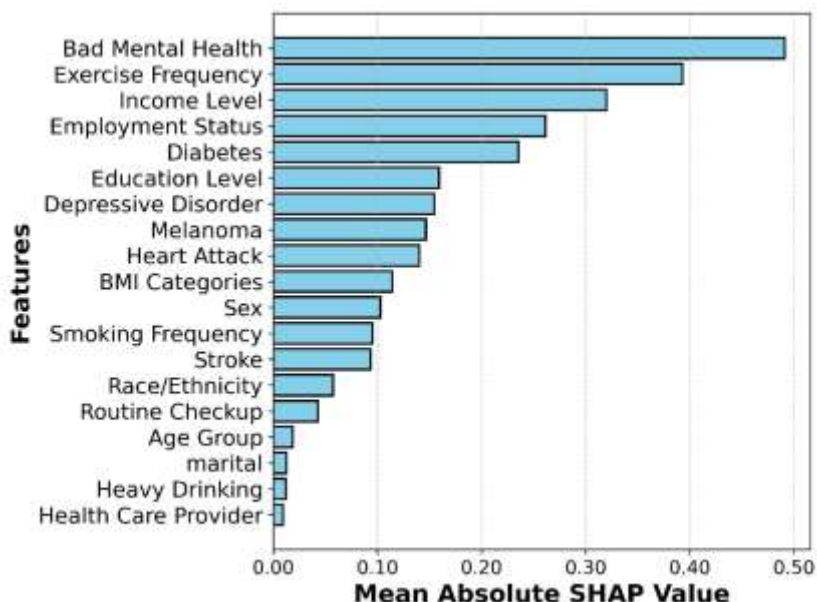


Figure 5: Overall feature importance ranking of predictive model variables.

### DISCUSSION

The findings indicate that socio-economic and behavioral factors play a significant role in shaping one’s perception of health. In particular, factors such as mental health status, level of income, employment status and physical exercise emerged as the strongest predictors of SRH in our models demonstrate that individuals assess their health in relation to their lived social experiences. This suggests that SRH is not solely determined by medical conditions but also influenced by broader social and lifestyle related factors (Hanmer, 2021; Nehme *et al.*, 2024; Safaei & Saliminezhad, 2023). The strong influence of socio-

economic status on SRH aligns with previous research (Kim *et al.*, 2023). Financial stability has a significant impact on health outcomes. Higher income provides access to a better living condition which contributes to a more positive perception of health. Similarly, employment status was also found to be a strong predictor which indicates economic security and stress significantly influence perceived health. Behavioral factors, particularly physical exercise and mental health notably influence SRH (Gumà-Lao & Arpino, 2023). Recent work on the vulnerable people also found that socio-economic status, education, and mental well-being significantly influenced perceived well-being among older adults which is consistent with our findings from the U.S. BRFSS dataset (Anwar *et al.*, 2024). Interestingly, traditional health risk behaviors such as heavy drinking, smoking frequency, BMI, routine checkup and healthcare access were not identified as strong predictors of SRH. Although these factors are well documented contributors for long term health results, their impact on SRH is low and that may suggest people do not think these are direct contributors towards their health and wellbeing. One possible explanation is that individuals may not directly link smoking and heavy alcohol consumption to their health unless severe health complications occur (Zhao *et al.*, 2018). Additionally, routine checkups and access to healthcare services did not significantly contribute to SRH which could suggest people understand health in the context of everyday life, rather than from clinical care (Zimmerman & Anderson, 2019).

These findings can be interpreted through the two theoretical lenses: the Social Determinants of Health (SDOH) framework and the Health Belief Model (HBM). SDOH posits that an individual's economic stability, education, access to social support, and living environment heavily influence their health outcomes and perceptions. This observation aligns with the understanding that health perception is a subjective construct that is shaped by life experience and social context rather than just clinical symptoms. Furthermore, HBM can be drawn on to explain why health risk behaviors such as alcohol consumption rate, smoking frequency, BMI, routine checkup and healthcare access were not identified as strong predictors of SRH. HBM suggests that individuals take health-related actions based on perceived susceptibility and severity rather than objective risks. For instance, individuals may not perceive smoking or alcohol consumption as immediately detrimental to their health unless they face severe complications. This indicates that people may not internalize risky behaviors as harmful until the effects become acute or life-threatening. These findings underscore the growing relevance of patient centric healthcare models that leverage patient's behavioral data for more sustainable and individualized care solutions (Zahid *et al.*, 2025; Zahid *et al.*, 2023; Zahid *et al.*, 2021).

Methodologically, the study also shows that applying a technique of Repeated Under-sampling is a viable approach for dealing with class imbalance in the BRFSS data for a subjective target variable like SRH. Traditional oversampling methods introduce synthetic samples, which may sometimes lead to overfitting. Repeated Under-sampling provided a balanced representation of both classes of SRH without compromising the model performance. Public health research that is subject to imbalanced datasets may find this technique useful. The applied ML models gave productive insights from SRH. Logistic Regression demonstrated strong performance with an AUC of 81% while the Stacked Model provided a slight improvement. The key predictors were confirmed with the application of Shapley Additive Explanations (SHAP), which provided a level of understanding and clarity with respect to the model's decision making and level of importance of each determinant. These findings highlight the potential of machine learning in uncovering the determinants of perceived health.

### LIMITATIONS AND FUTURE WORK

Despite the strengths of this study, the study draws some limitations which are potential for further research. This is a cross-sectional study, it captures the health perceptions at a single point of time. So, it is not possible to determine the causal relationships between the predictors and SRH. In order to understand how these components affect SRH over a longer period and whether the variations improve or worsen the person's perception towards their health, longitudinal studies need to be conducted (Rähse *et al.*, 2025). Additionally prior work has shown, trust is a critical barrier in large-scale digital health initiatives so integrating transparency and user-centered design may improve future adoption and engagement (Sharma *et al.*, 2020; Zahid *et al.*, 2025; Zahid *et al.*, 2023; Zahid *et al.*, 2024). Additionally, qualitative research could complement these quantitative findings by examining how people interpret and articulate their sense of health. For example, a study on stroke survivors have provided valuable insights into how people experience and rate their health subjectively (Mavaddat *et al.*, 2018).

### CONCLUSION

In this study of BRFSS 2023 data, interpretable machine learning identified socio-economic and behavioral factors, especially mental health, physical activity, income, and employment as primary determinants of self-rated health. Traditional clinical risk markers (e.g., smoking, heavy drinking, routine checkups) showed limited influence on predicting perceived health. These findings, interpreted via SDOH and HBM frameworks, imply that policies aiming to improve perceived population health should prioritize social and behavioral determinants alongside clinical care. Methodologically, repeated under-sampling provided robust class balance without synthetic data generation and can be a practical tool for similar survey-based prediction tasks. Future work should evaluate temporal dynamics and subgroup validity and release reproducible code to support replication.

### ACKNOWLEDGMENTS

We gratefully acknowledge the participants of the Behavioral Risk Factor Surveillance System (BRFSS) survey for their contributions, without whom this research would not have been possible. We also thank the Centers for Disease Control and Prevention (CDC) for making the BRFSS dataset available for research.

## AUTHOR CONTRIBUTIONS

All authors have contributed equally to the work.

## FUNDING

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

## CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

## DATA AVAILABILITY

The BRFSS dataset used in this study is publicly available at CDC BRFSS website - <https://www.cdc.gov/brfss>. The code will be provided upon reasonable request.

## REFERENCES

- Alejo, R., Sotoca, J.M., Valdovinos, R.M., & Toribio, P. (2010). Edited nearest neighbor rule for improving neural networks classifications. In L. Zhang, B.-L. Lu, & J. Kwok (Eds.), *Advances in Neural Networks—ISNN 2010* (Vol. 6063, pp. 303–310). Berlin, Germany: Springer. [https://doi.org/10.1007/978-3-642-13278-0\\_39](https://doi.org/10.1007/978-3-642-13278-0_39)
- Alyafei, A., & Easton-Carr, R. (2025). The health belief model of behavior change. In *StatPearls*. StatPearls Publishing. <http://www.ncbi.nlm.nih.gov/books/NBK606120/>
- Anwar, A., Akter, N., Yadav, U.N., Ghimire, S., Bhattacharjee, S., Eusufzai, S.Z., ... Mistry, S.K. (2024). Assessment of mental well-being and its socio-economic determinants among older adults in the Rohingya refugee camp of Bangladesh. *Scientific Reports*, *14*(1), 17918. <https://doi.org/10.1038/s41598-024-68795-9>
- Balaj, M. (2022). Self-reported health and the social body. *Social Theory & Health*, *20*(1), 71–89. <https://doi.org/10.1057/s41285-020-00150-0>
- Barreto, S.M., & Figueiredo, R.C.D. (2009). Doença crônica, auto-avaliação de saúde e comportamento de risco: Diferença de gênero. *Revista de Saúde Pública*, *43*(suppl 2), 38–47. <https://doi.org/10.1590/S0034-89102009000900006>
- Centers for Disease Control and Prevention. (n.d.). *Behavioral Risk Factor Surveillance System [Dataset]*. Retrieved from [https://www.cdc.gov/brfss/annual\\_data/annual\\_2023.html](https://www.cdc.gov/brfss/annual_data/annual_2023.html) (accessed 25 October 2024).
- Chaturvedi, A., Zhu, A., Gadela, N.V., Prabhakaran, D., & Jafar, T.H. (2024). Social determinants of health and disparities in hypertension and cardiovascular Diseases. *Hypertension*, *81*(3), 387–399. <https://doi.org/10.1161/HYPERTENSIONAHA.123.21354>
- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. <https://doi.org/10.1613/jair.953>
- Chen, S., Yu, J., Chamouni, S., Wang, Y., & Li, Y. (2024). Integrating machine learning and artificial intelligence in life-course epidemiology: Pathways to innovative public health solutions. *BMC Medicine*, *22*(1), 354. <https://doi.org/10.1186/s12916-024-03566-x>
- Choi, J.H., & Jung, D.H. (2025). The role of psychological factors in predicting self-rated health: Implications from machine learning models. *Psychology, Health & Medicine*, *30*(6), 1158–1170. <https://doi.org/10.1080/13548506.2025.2450546>
- Chowdhury, M.M., Ayon, R.S., & Hossain, M.S. (2024). An investigation of machine learning algorithms and data augmentation techniques for diabetes diagnosis using class imbalanced BRFSS dataset. *Healthcare Analytics*, *5*, 100297. <https://doi.org/10.1016/j.health.2023.100297>
- Dierx, J.A.J., & Kasper, H.D.P. (2022). The magnitude and importance of perceived health dimensions define effective tailor-made health-promoting interventions per targeted socioeconomic group. *Frontiers in Public Health*, *10*, 849013. <https://doi.org/10.3389/fpubh.2022.849013>
- El Morr, C., Jammal, M., Bou-Hamad, I., Hijazi, S., Ayna, D., Romani, M., & Hoteit, R. (2024). Predictive machine learning models for assessing Lebanese university students' depression, anxiety, and stress during COVID-19. *Journal of Primary Care & Community Health*, *15*, 21501319241235588. <https://doi.org/10.1177/21501319241235588>
- Fregoso-Aparicio, L., Noguez, J., Montesinos, L., & García-García, J. A. (2021). Machine learning and deep learning predictive models for type 2 diabetes: A systematic review. *Diabetology & Metabolic Syndrome*, *13*(1), 148. <https://doi.org/10.1186/s13098-021-00767-9>
- Google. (n.d.). *Google colabatory* [Computer software]. <https://colab.research.google.com>
- Gumà-Lao, J., & Arpino, B. (2023). A machine learning approach to determine the influence of specific health conditions on self-rated health across education groups. *BMC Public Health*, *23*(1), 131. <https://doi.org/10.1186/s12889-023-15053-8>
- Gündoğdu, S. (2023). Efficient prediction of early-stage diabetes using XGBoost classifier with random forest feature selection technique. *Multimedia Tools and Applications*, *82*(22), 34163–34181. <https://doi.org/10.1007/s11042-023-15165-8>
- He, H., & Garcia, E.A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263–1284. <https://doi.org/10.1109/TKDE.2008.239>
- Halder, R.K., Uddin, M.N., Uddin, Md. A., Aryal, S., & Khraisat, A. (2024). Enhancing k-nearest neighbor algorithm: A comprehensive review and performance analysis of modifications. *Journal of Big Data*, *11*(1), 113. <https://doi.org/10.1186/s40537-024-00973-y>
- Hamplová, D., Klusáček, J., & Mráček, T. (2022). Assessment of self-rated health: The relative importance of physiological, mental, and socioeconomic factors. *Plos One*, *17*(4), e0267115. <https://doi.org/10.1371/journal.pone.0267115>

- Hanmer, J. (2021). Measuring population health: Association of self-rated health and PROMIS measures with social determinants of health in a cross-sectional survey of the US population. *Health and Quality of Life Outcomes*, 19(1), 221. <https://doi.org/10.1186/s12955-021-01854-1>
- Idler, E.L., & Benyamini, Y. (1997). Self-rated health and mortality: A review of twenty-seven community studies. *Journal of Health and Social Behavior*, 38(1), 21–37.
- Imoisili, O. E., Chung, A., Tong, X., Hayes, D. K., & Loustalot, F. (2024). Prevalence of stroke—Behavioral risk factor surveillance system, United States, 2011–2022. *MMWR. Morbidity and Mortality Weekly Report*, 73(20), 449–455. <https://doi.org/10.15585/mmwr.mm7320a1>
- Janz, N.K., & Becker, M.H. (1984). The health belief model: A decade later. *Health Education Quarterly*, 11(1), 1–47. <https://doi.org/10.1177/109019818401100101>
- Joloudari, J.H., Marefat, A., Nematollahi, M.A., Oyelere, S.S., & Hussain, S. (2023). Effective class-imbalance learning based on SMOTE and convolutional neural networks. *Applied Sciences*, 13(6), 4006. <https://doi.org/10.3390/app13064006>
- Kim, Y., Vazquez, C., & Cubbin, C. (2023). Socioeconomic disparities in health outcomes in the United States in the late 2010s: Results from four national population-based studies. *Archives of Public Health*, 81(1), 15. <https://doi.org/10.1186/s13690-023-01026-1>
- Ku, W.L., & Min, H. (2024). Evaluating machine learning stability in predicting depression and anxiety amidst subjective response errors. *Healthcare*, 12(6), 625. <https://doi.org/10.3390/healthcare12060625>
- Latham, K., & Peek, C.W. (2013). Self-rated health and morbidity onset among late midlife U.S. adults. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, 68(1), 107–116. <https://doi.org/10.1093/geronb/gbs104>
- Limbu, Y.B., Gautam, R.K., & Pham, L. (2022). The health belief model applied to COVID-19 vaccine hesitancy: A Systematic Review. *Vaccines*, 10(6), 973. <https://doi.org/10.3390/vaccines10060973>
- Linton, M.J., Dieppe, P., & Medina-Lara, A. (2016). Review of 99 self-report measures for assessing well-being in adults: Exploring dimensions of well-being and developments over time. *BMJ Open*, 6(7), e010641. <https://doi.org/10.1136/bmjopen-2015-010641>
- Lozem, G., Cook, S., Leon, D.A., Emaus, N., & Schirmer, H. (2020). Self-reported health as a predictor of mortality: A cohort study of its relation to other health measurements and observation time. *Scientific Reports*, 10(1), 4886. <https://doi.org/10.1038/s41598-020-61603-0>
- Lundberg, S., & Lee, S.I. (2017). *A unified approach to interpreting model predictions (Version 2)*. arXiv. <https://doi.org/10.48550/ARXIV.1705.07874>
- Mavaddat, N., Sadler, E., Lim, L., Williams, K., Warburton, E., Kinmonth, A. L., Mant, J., Burt, J., & McKeivitt, C. (2018). Perceptions of self-rated health among stroke survivors: A qualitative study in the United Kingdom. *BMC Geriatrics*, 18(1), 81. <https://doi.org/10.1186/s12877-018-0765-8>
- Mu, F., Liu, J., Lou, H., Zhu, W., Wang, Z., & Li, B. (2024). Influence of physical exercise on negative emotions in college students: Chain mediating role of sleep quality and self-rated health. *Frontiers in Public Health*, 12, 1402801. <https://doi.org/10.3389/fpubh.2024.1402801>
- Nehme, M., Schrepft, S., Baysson, H., Pullen, N., Rouzinov, S., Stringhini, S., ... Guessous, I. (2024). Associations between healthy behaviors and persistently favorable self-rated health in a longitudinal population-based study in Switzerland. *Journal of General Internal Medicine*, 39(10), 1828–1838. <https://doi.org/10.1007/s11606-024-08739-1>
- Nelson, D.E., Holtzman, D., Bolen, J., Stanwyck, C.A., & Mack, K.A. (2001). Reliability and validity of measures from the Behavioral Risk Factor Surveillance System (BRFSS). *Sozial-Und Praventivmedizin*, 46 Suppl 1, S3-42.
- Olisah, C.C., Smith, L., & Smith, M. (2022). Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective. *Computer Methods and Programs in Biomedicine*, 220, 106773. <https://doi.org/10.1016/j.cmpb.2022.106773>
- Ozckakir, A., Oflu Dogan, F., Cakir, Y.T., Bayram, N., & Bilgel, N. (2014). Subjective well-being among primary health care patients. *Plos One*, 9(12), e114496. <https://doi.org/10.1371/journal.pone.0114496>
- Parga, J.P.F.A., Tiznado-Aitken, I., Jamal, S., Farber, S., Yu, A., & Higgins, C. (2024). Perceived accessibility and self-rated health: Examining subjective well-being in the suburbs of Scarborough, Canada. *Transportation Research Part A: Policy and Practice*, 190, 104261. <https://doi.org/10.1016/j.tra.2024.104261>
- Pierannunzi, C., Hu, S.S., & Balluz, L. (2013). A systematic review of publications assessing reliability and validity of the Behavioral Risk Factor Surveillance System (BRFSS), 2004–2011. *BMC Medical Research Methodology*, 13(1), 49. <https://doi.org/10.1186/1471-2288-13-49>
- Rähse, T., Richter, M., & Knöchelmann, A. (2025). Continuing exposure to disadvantageous material and perceived economic factors on self-rated health in different life stages: Fixed effects analyses with data from the German Socioeconomic Panel. *BMC Public Health*, 25(1), 446. <https://doi.org/10.1186/s12889-024-21135-y>
- Sabouri, Z., Gherabi, N., Nasri, M., Amnai, M., Massari, H. E., & Moustati, I. (2023). Prediction of depression via supervised learning models: Performance comparison and analysis. *International Journal of Online & Biomedical Engineering*, 19(9), 93–107. <https://doi.org/10.3991/ijoe.v19i09.39823>
- Safaei, J., & Saliminezhad, A. (2023). Health impacts of social determinants and lifestyle behaviours: Some evidence from Canadian provinces. *Health Promotion International*, 38(4), daad067. <https://doi.org/10.1093/heapro/daad067>
- Sekkarie, A., Fang, J., Hayes, D., & Loustalot, F. (2024). Prevalence of self-reported hypertension and antihypertensive medication use among adults—United States, 2017–2021. *MMWR. Morbidity and Mortality Weekly Report*, 73(9), 191–198. <https://doi.org/10.15585/mmwr.mm7309a1>

- Sharma, R., Zhang, C., Wingreen, S.C., Kshetri, N., & Zahid, A. (2020). Design of blockchain-based precision health-care using soft systems methodology. *Industrial Management & Data Systems*, 120(3), 608-632.
- Subramanian, S., Huijts, T., & Avendano, M. (2010). Self-reported health assessments in the 2002 World Health Survey: How do they correlate with education? *Bulletin of the World Health Organization*, 88(2), 131-138. <https://doi.org/10.2471/BLT.09.067058>
- Suess, C., Maddock, J.E., Dogru, T., Mody, M., & Lee, S. (2022). Using the Health Belief Model to examine travelers' willingness to vaccinate and support for vaccination requirements prior to travel. *Tourism Management*, 88, 104405. <https://doi.org/10.1016/j.tourman.2021.104405>
- Tadiri, C.P., Gisinger, T., Kautzky-Willer, A., Kublickiene, K., Herrero, M.T., Norris, C.M., ...Pilote, L. (2021). Determinants of perceived health and unmet healthcare needs in universal healthcare systems with high gender equality. *BMC Public Health*, 21(1), 1488. <https://doi.org/10.1186/s12889-021-11531-z>
- World Health Organization. (2004). *The World Health Organization quality of life (WHOQOL)—BREF, 2012 revision*. World Health Organization. Retrieved from <https://iris.who.int/handle/10665/77773> (accessed 15 July 2025).
- Wang, H., Liang, Q., Hancock, J.T., & Khoshgoftaar, T.M. (2024). Feature selection strategies: A comparative analysis of SHAP-value and importance-based methods. *Journal of Big Data*, 11(1), 44. <https://doi.org/10.1186/s40537-024-00905-w>
- Weech-Maldonado, R., Miller, M.J., & Lord, J.C. (2017). The relationships among socio-demographics, perceived health, and happiness. *Applied Research in Quality of Life*, 12(2), 289-302. <https://doi.org/10.1007/s11482-017-9517-8>
- Wongvorachan, T., He, S., & Bulut, O. (2023). A comparison of undersampling, oversampling, and SMOTE methods for dealing with imbalanced classification in educational data mining. *Information*, 14(1), 54. <https://doi.org/10.3390/info14010054>
- World Health Organization. (2024). *Operational framework for monitoring social determinants of health equity*. World Health Organization.
- Wuorela, M., Lavonius, S., Salminen, M., Vahlberg, T., Viitanen, M., & Viikari, L. (2020). Self-rated health and objective health status as predictors of all-cause mortality among older people: A prospective study with a 5-, 10-, and 27-year follow-up. *BMC Geriatrics*, 20(1), 120. <https://doi.org/10.1186/s12877-020-01516-9>
- Liu, X.Y., Wu, J., & Zhou, Z.H. (2009). Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(2), 539-550. <https://doi.org/10.1109/TSMCB.2008.2007853>
- Zahid, A., Sharma, R., & Ahmed, R. (2025). Blockchain for building trust in genomic initiatives. *World Medical & Health Policy*, 17(1), 176-187. <https://doi.org/10.1002/wmh3.641>
- Zahid, A., Wingreen, S., & Sharma, R. (2025). Affordance theory as a conceptual foundation to design Precision Healthcare ecosystems. In *Research Handbook on Health Information Systems* (pp. 31-44). Cheltenham, UK: Edward Elgar Publishing.
- Zahid, A., & Sharma, R. (2023). Personalized health care in a data-driven era: A post-COVID-19 retrospective. *Mayo Clinic Proceedings: Digital Health*, 1(2), 162-171.
- Zahid, A., Yadav, U.N., & Mistry, S.K. (2023). Digital health services in Bangladesh—The need for a sustainable design framework. *World Medical & Health Policy*, 15(4), 682-695.
- Zahid, A. (2021). *An overview of emerging digital health informatics\**. CRC Press. <https://doi.org/10.1201/9781003224150>
- Zahid, A., Wingreen, S., Sharma, R.S., & Inthiran, A. (2024). Digital sustainable healthcare: Field validation of a patient-centric use-case. *PACIS 2024 Proceedings*, 5. Retrieved from [https://aisel.aisnet.org/pacis2024/track11\\_healthit/track11\\_healthit/5](https://aisel.aisnet.org/pacis2024/track11_healthit/track11_healthit/5) (access 15 July 2025).
- Zaman, S.M., Qureshi, W.M., Raihan, M.M.S., Monjur, O., & Shams, A.B. (2021). *Survival prediction of heart failure patients using stacked ensemble machine learning algorithm (Version 1)*. arXiv. <https://doi.org/10.48550/ARXIV.2108.13367>
- Zhao, G., Okoro, C.A., Hsia, J., & Town, M. (2018). Self-perceived poor/fair health, frequent mental distress, and health insurance status among working-aged US adults. *Preventing Chronic Disease*, 15, 170523. <https://doi.org/10.5888/pcd15.170523>
- Zimmerman, F.J., & Anderson, N.W. (2019). Trends in health equity in the United States by race/ethnicity, sex, and income, 1993-2017. *JAMA Network Open*, 2(6), e196386. <https://doi.org/10.1001/jamanetworkopen.2019.6386>
- Zuo, D., Yang, L., Jin, Y., Qi, H., Liu, Y., & Ren, L. (2023). Machine learning-based models for the prediction of breast cancer recurrence risk. *BMC Medical Informatics and Decision Making*, 23(1), 276. <https://doi.org/10.1186/s12911-023-02377-z>