Identifying individuals at risk of post-stroke depression

Development and validation of a predictive model

Saeed A. Alqahtani, MD, PhD.

ABSTRACT

الأهداف: تحديد العوامل المرتبطة بالاكتئاب ما بعد السكتة الدماغية وتطوير نموذج تنبؤي للتعلم الآلي باستخدام مجموعة بيانات كبيرة، مع مراعاة العوامل الاجتماعية والديموغرافية ونمط الحياة والعوامل السريرية.

المنهجية: استخدمت دراستنا لعام 2025 بيانات من BREFSS لعام 2023، والتي تم إصدارها في سبتمبر 2024. وقد تم إجراء معالجة البيانات في Google Colab باستخدام لغة Python. قمنا بإجراء إحصائيات وصفية، وتحليل الانحدار اللوجستي، وتحليل أهمية الميزات (المعلومات المتبادلة والمعلومات المتبادلة المعدلة). تم تدريب وتقييم 4 نماذج للتعلم الآلي: random forest, decision tree, gradient boosting و logistic regression. تقييم أداء النموذج باستخدام accuracy, precision recall, F1-score و AUC-ROC. قمنا بضبط النموذج الأفضل أداءً باستخدام GridSearchCV مع التحقق المتقاطع خماسي الطيات (5-fold cross-validation).

النتائج: ارتبط التقدم في العمر، وكون الشخص ذكرًا، والحالة الزوجية (متزوج)، وارتفاع الدخل، والنشاط البدني بانخفاض احتمالات الإصابة بالاكتئاب ما بعد السكتة الدماغية. وارتبطت السمنة والتدخين ومرض السكري وارتفاع الكوليسترول بزيادة احتمالات الإصابة بالاكتئاب ما بعد السكتة الدماغية. كان العمر والجنس أكثر السمات إفادة للتنبؤ بالاكتئاب ما بعد السكتة الدماغية. أظهر نموذج random forest أفضل أداء في التنبؤ بالاكتئاب ما بعد السكتة الدماغية (accuracy=0.73, precision=0.71 recall=0.77, F1-score=0.74 و AUC-ROC=0.81)، وقد تحسن الأداء بشكل أكبر من خلال تحسين المعلمات.

الخلاصة: ينطوي سبب الاكتئاب ما بعد السكتة الدماغية على تداخل معقد بين العوامل الاجتماعية والديموغرافية ونمط الحياة والعوامل السريرية، ولا سيما العمر والجنس. ويتنبأ نموذج الغابة العشوائية بشكل فعال بالاكتئاب ما بعد السكتة الدماغية، مما يسلط الضوء على الحاجة إلى التقييم الشامل، والتدخل المبكر، وإدارة عوامل الخطر القابلة للتعديل (السمنة، والتدخين، والخمول) لتحسين نتائج الناجين من السكتة الدماغية.

Objectives: To identify the factors associated with post-stroke depression (PSD) and develop a machine learning predictive model using a large dataset, considering sociodemographic, lifestyle, and clinical factors.

Methods: Our 2025 study used data from the 2023 Behavioral Risk Factor Surveillance System, released in September 2024. Data processing was carried out using Google Colab and Python. We carried out descriptive statistics, logistic regression, and feature importance analyses (mutual information and adjusted mutual information). A total of 4 machine-learning models were trained and evaluated: random forest, decision tree, gradient boosting, and logistic regression. Model performance was assessed using the accuracy, precision, recall, harmonic mean of precision and recall (F1-score), and area under the curve - receiver operating characteristic (AUC-ROC). The best-performing model was fine-tuned using GridSearchCV with 5-fold cross-validation.

Results: Increasing age, male gender, being married, higher income, and physical activity were associated with lower odds of PSD. Obesity, smoking, diabetes, and high cholesterol are associated with increased odds of PSD. Age and gender were the most informative features for predicting the PSD. Random forest demonstrated the best performance for predicting PSD (accuracy=0.73, precision=0.71, recall=0.77, F1-score=0.74, and AUC-ROC=0.81), which was further improved by hyperparameter optimization.

Conclusion: Post-stroke depression's complex etiology involves sociodemographic, lifestyle, and clinical factors, notably age and gender. A random forest model effectively predicts PSD, highlighting the need for comprehensive assessment, early intervention, and management of modifiable risks (obesity, smoking, and inactivity) to improve stroke survivors' outcomes.

Keywords: post-stroke depression, risk factors, machine learning, mutual information, logistic regression

Saudi Med J 2025; Vol. 46 (5): 497-506
doi: 10.15537/smj.2025.46.5.20250080

From the Department of Basic Medical Sciences, Taibah University, Al-Madinah Al-Munawarah, Kingdom of Saudi Arabia.

Received 21st January 2025. Accepted 4th April 2025.

Address correspondence and reprint request to: Dr. Saeed A. Alqahtani, Department of Basic Medical Sciences, Taibah University, Al-Madinah Al-Munawarah, Kingdom of Saudi Arabia.
E-mail: samqahtani@taibahu.edu.sa
ORCID ID: https://orcid.org/0000-0002-2419-2188

Stroke is a leading cause of disability and mortality worldwide, imposing a significant burden on individuals, families, and healthcare system. Among the most prevalent and debilitating post-stroke complications is post-stroke depression (PSD). It is a prevalent and debilitating condition affecting approximately one-third of stroke survivors.[1] Post-stroke depression significantly impacts recovery, and quality of life, and increases mortality among stroke survivors.[2] The impact of PSD extends beyond the individual's emotional well-being. It is associated with poorer functional recovery, increased disability, higher healthcare utilization, reduced quality of life, and even increased mortality.[3] Given the profound impact of PSD on stroke recovery and outcomes, early identification and effective management of this condition are of paramount importance.

Recent research has identified several sociodemographic factors associated with PSD, highlighting the complex and multifactorial nature of this condition. Age is a critical factor, as older adults are more susceptible to moderate depressive symptoms post-stroke.[4] Gender differences are also evident, with women more likely to experience PSD than men.[4] A history of mental disorders, particularly previous depressive episodes, significantly increases the risk of PSD, underscoring the importance of mental health history in predicting PSD.[5] Social support emerges as a protective factor, with higher perceived social support correlating with lower PSD risk.[5] Additionally, cultural context influences PSD prevalence, as seen in variations between countries such as Australia, New Zealand, and Vietnam.[6]

Lifestyle factors, such as obesity, physical inactivity, and smoking have been identified as significant modifiable risk factors for PSD. Obesity and physical inactivity are linked to increased risk of stroke, and by extension, they contribute to the risk of PSD due to their role in exacerbating cardiovascular and metabolic conditions that can lead to stroke.[7] Smoking, another critical lifestyle factor, is associated with both the incidence of stroke and the development of depression, including PSD.[8] The interplay between these lifestyle factors and PSD is complex, as they not only contribute to the initial risk of stroke but also affect the psychological and physical recovery post-stroke.[9] For instance, smoking has been directly linked to increased depressive symptoms, while physical inactivity can exacerbate cognitive and functional impairments, further increasing the risk of PSD.[8] Moreover, lifestyle modifications such as increased physical activity, smoking cessation, and maintaining a healthy weight are emphasized as effective secondary prevention measures to reduce the risk of stroke recurrence and associated depressive symptoms.[10]

The presence of comorbid conditions such as anxiety disorders, previous depression, and other psychiatric disorders notably increases the likelihood of developing PSD. For instance, patients with a comorbid anxiety disorder are 5.9 times more likely to experience PSD, and those with a history of depression treatment before stroke are 3.0 times more likely to develop PSD.[11] Additionally, medical conditions like diabetes and cardiovascular disease are a potential direct pathophysiological links to depressive disorders, which can exacerbate the risk of PSD.[12] The bidirectional relationship between cerebrovascular diseases and depression further complicates this scenario, as depression itself can increase the risk of stroke, creating a cyclical risk factor.[13] Moreover, the severity of functional impairment post-stroke, often compounded by comorbid conditions, is associated with a higher risk of PSD, suggesting that both psychological and neurobiological factors play a role in its development.[5,14] These findings underscore the importance of comprehensive post-stroke care that addresses not only the physical rehabilitation but also the management of psychiatric comorbidities to improve overall outcomes for stroke survivors.[15]

Predicting PSD involves understanding the complex interplay of risk factors impacting stroke survivors' recovery and well-being. Key sociodemographic factors include age, gender, marital status, and income, with the impact of education often intertwined with socioeconomic status.[16-24] Lifestyle choices significantly contribute, as obesity and smoking increase PSD risk, while physical activity offers protection.[20,25-33] Certain comorbidities, specifically diabetes and high cholesterol, independently elevate depression risk.[28,34-36] Comprehensive risk assessment, encompassing these diverse factors, is thus crucial for accurate PSD prediction and targeted interventions. Therefore, implementation of predictive models in clinical practice can lead to significant improvements in patient care and outcomes. By identifying high-risk patients early, clinicians can allocate resources more effectively, reducing the overall burden on the healthcare systems. Additionally, early intervention can lead to better functional recovery, improved quality of life, and reduced mortality rates among stroke survivors.

Disclosure. Author has no conflict of interests, and the work was not supported or funded by any drug company.

This study aimed to identify factors associated with depression in stroke patients and to develop a machine learning predictive model for PSD using a large dataset, taking into account a wide range of sociodemographic, lifestyle, and clinical factors.

Methods. Our 2025 study used publicly available data from the 2023 Behavioral Risk Factor Surveillance System (BRFSS) to examine health behaviors in the United States. The most recent version of this dataset was released in September 2024. The BRFSS is overseen by the US Centers for Disease Control and Prevention (CDC) released under the CC0 1.0 Universal Public Domain Dedication license, thus eliminating the need for ethical approval or informed consent for its use.

Data processing was carried out in a Google Colaboratory (Colab) environment using Python. These include data cleaning, feature selection, and feature engineering. The initial dataset encompassed 433,323 records. This was filtered to retain 14 categorical variables deemed relevant to stroke and depression. To examine PSD, a new dataset was created from the BRFSS, comprising only respondents who reported ever experiencing a stroke (approximately 4.21% of the total BRFSS sample). This stroke-specific dataset was used in all subsequent analyses. A binary outcome variable, depression stroke, was constructed to represent PSD by combining the depression status and stroke status variables. Since the dataset was already limited to stroke respondents, the stroke status variable was, by definition, complete for all individuals in this new dataset. The depression stroke variable required complete data on depression status. Therefore, records with missing data on depression status was excluded from the final analysis. This data cleaning process resulted in a final analytical dataset of 17,460 individuals. The missingness rate for depression status within this stroke-specific dataset was approximately 0.9 %. Missing values in other variables were imputed using logistic regression. Specifically, a logistic regression model was trained using the complete cases to predict the missing values in the incomplete cases. The predicted probabilities were then used to assign the most likely values to the missing entries.

To quantify the strength and nature of the association between individual predictors and PSD, we carried out mutual information (MI) and adjusted mutual information (AMI) analyses. Mutual information, a concept from information theory, measures the amount of information one random variable reveals regarding another. In this context, a higher MI value between a predictor and PSD suggests that knowing the value of the predictor significantly reduces uncertainty regarding the presence of PSD. Unlike correlations, MI can capture non-linear relationships. Adjusted mutual information builds upon MI by adjusting the score to account for chance agreement. This correction is crucial, as MI can be artificially inflated for variables with many different values, even if their association is weak. Therefore, AMI provides a more reliable measure of the true association, particularly when comparing predictors with varying numbers of levels. Mutual information scores were calculated using the \`mutual_info_classif\` function from the \`sklearn.feature_selection\` module, which is specifically designed for classification tasks. Adjusted mutual information scores were derived using the \`adjusted_mutual_info_score\` function from the \`sklearn.metrics\` module. To facilitate interpretability and comparison across predictors, both MI and AMI scores were normalized to percentages, representing the relative contribution of each predictor to the total information gain regarding PSD. These percentages reflect the relative importance of each predictor in a univariate context, independent of any predictive model.

To predict the likelihood of an individual with a stroke history reporting a depression diagnosis, we employed and compared 4 machine learning models: random forest classifier, decision tree classifier, gradient boosting classifier, and logistic regression. Recognizing the "ever diagnosed" nature of the BRFSS data, which captures population-level patterns of reported diagnoses, we focused on predicting the probability of a depression diagnosis rather than diagnosing current PSD. Pre-existing depression was explicitly accounted for as an input variable, acknowledging the complex, bidirectional relationship between stroke and depression. To address class imbalance within the dataset, the synthetic minority over-sampling technique was applied. Model performance was rigorously evaluated using several metrics: accuracy, precision, recall, harmonic mean of precision and recall (F1-score), and area under the curve - receiver operating characteristic (AUC-ROC). The model exhibiting superior performance underwent hyperparameter optimization using GridSearchCV with 5-fold cross-validation to ensure robustness of the optimized model's performance.

Statistical analysis. Descriptive statistics were generated for all 15 categorical variables (including the newly created "depression stroke" variable) prior to any data preprocessing. This provided an overview of the variable distributions within the dataset. Data analysis was carried out using Python's Pandas library where frequency and percentage distributions of each category were computed, providing an overview of the dataset's composition. Further, all the subsequent analyses were carried out using the created stroke-specific dataset.

We carried out logistic regression to analyze the relationships between the predictor variables and the target variable (depression stroke). Odds ratios (ORs), 95% confidence intervals (CIs), and *p*-values were calculated to assess the strength of these associations, considering *p*-values below 0.05 as significant. For variables with more than 2 categories, group 1 was used as the reference group, except for the body mass index (BMI) category variable, where group 2 (normal weight) served as the reference.

Results. The sample (N=419,476) was relatively balanced in terms of gender (53% female). A small majority were married (52%) and reported alcohol consumption (56%). Most participants were physically active (75%) and non-smokers (89%). Regarding health conditions, most participants did not have diabetes (86%), hypertension (59%), or a history of stroke (96%). A minority reported elevated cholesterol (42%) or a depressive disorder (20%). Among those with a history of stroke, 30% were experiencing depression. The most frequent age groups were 65-69 (11%) years and 70-74 (10%) years. The most common education level was "graduated from college or technical school" (43%). The most frequent income category was $50,000 to <$100,000 (31%), and the most frequent BMI category was overweight (36%, Table 1).

A multiple logistic regression analysis was carried out to identify factors associated with PSD. Table 2 displays the ORs, 95% CIs, and *p*-values for each independent variable. Male gender was significantly associated with lower odds of PSD (OR=0.57, 95% CI: [0.53-0.61], *p*<0.05), as was increasing age. Compared to the youngest age group (18-24 years), the oldest group (80 years or older) had 84% lower odds of PSD (OR=0.16, 95% CI: [0.11-0.24], *p*<0.05). Being married was also associated with lower odds of PSD (OR=0.86, 95% CI: [0.80-0.93], *p*<0.05). Higher income was significantly associated with lower odds of PSD. Compared to the lowest income category, the highest income category had the lowest odds (OR=0.4, 95% CI: [0.30-0.54], *p*<0.05). Education level was not significantly associated with PSD. Obesity was significantly associated with increased odds of PSD in stroke patients (OR=1.33, 95% CI: [1.21-1.46], *p*<0.05). However, being underweight or overweight was not significantly associated with PSD. Physical activity was associated with lower odds of PSD (OR=0.77, 95% CI: [0.72-0.83], *p*<0.05), while a history of smoking was associated with higher

Table 1 - Descriptive analysis of the variables.

Variables	n (%)
Depression stroke	
No	12231 (0.1)
Yes	5229 (29.9)
Gender	
Female	220428 (52.5)
Male	199048 (47.5)
Marital status	
Not married	200055 (48.2)
Married	215236 (51.8)
Alcohol status	
Do not drink	184234 (43.9)
Yes	235242 (56.1)
Physical activity	
No physical activity or exercise in last 30 days	102760 (24.6)
Had physical activity or exercise	315502 (75.4)
Smoking status	
No	353604 (89.1)
Yes	43441 (10.9)
Diabetes status	
Not diabetic	358706 (85.7)
Diabetic	59786 (14.3)
Hypertension status	
No hypertension	248435 (59.5)
Have hypertension	169195 (40.5)
Cholesterol status	
Normal	213522 (58.4)
High	152114 (41.6)
Stroke status	
No	400461 (95.8)
Yes	17609 (4.2)
Depressive disorder	
No	332667 (79.8)
Yes	84333 (20.2)
Age groups (years)	
65-69	44491 (10.6)
70-74	42049 (10.0)
60-64	40438 (9.6)
80-older	37723 (9.0)
75-79	33407 (8.0)
55-59	32968 (7.9)
50-54	30131 (7.2)
40-44	27445 (6.5)
35-39	26188 (6.2)
45-49	26172 (6.2)
18-24	25962 (6.2)
30-34	24101 (5.7)
25-29	20804 (5.0)
Education level	
Graduated from college or technical school	179801 (42.9)
Attended college or technical school	110398 (26.3)
Graduated high school	103086 (24.6)
Did not graduate high school	23920 (5.7)
Income category	
$50,000 to <$100,000	103861 (30.9)
$100,000 to <$200,000	74632 (22.2)
$35,000 to <$50,000	45827 (13.7)
$25,000 to <$35,000	37045 (11.0)
$15,000 to <$25,000	29749 (8.9)
$200,000 or more	26223 (7.8)
Less than $15,000	18270 (5.4)
BMI category	
Overweight	135644 (35.7)
Obese	123934 (32.6)
Normal weight	114035 (30.0)
Underweight	6638 (1.7)

Values are presented as numbers and percentages (%).
BMI: body mass index

Table 2 - Logistic regression analysis.

Variables	Odds ratio	95% CI	*P*-values
Intercept	1.86	1.21-2.85	<0.05
BMI			
Obese	1.33	1.21-1.46	<0.05
Overweight	1.06	0.97-1.16	0.23
Underweight	1.16	0.90-1.49	0.25
Gender	0.57	0.53-0.61	<0.05
Marital status	0.86	0.80-0.93	<0.05
Physical activity	0.77	0.72-0.83	<0.05
Smoking status	1.5	1.37-1.65	<0.05
Diabetes status	1.22	1.13-1.31	<0.05
Hypertension status	0.96	0.88-1.04	0.3
Cholesterol status	1.4	1.30-1.51	<0.05
Alcohol status	0.97	0.90-1.05	0.43
Age (years)			
25-29	1.02	0.59-1.78	0.94
30-34	0.74	0.45-1.24	0.25
35-39	0.94	0.59-1.51	0.81
40-44	0.78	0.50-1.22	0.28
45-49	0.84	0.55-1.30	0.44
50-54	0.58	0.38-0.89	<0.05
55-59	0.58	0.38-0.88	<0.05
60-64	0.51	0.34-0.76	<0.05
65-69	0.36	0.24-0.54	<0.05
70-74	0.32	0.21-0.47	<0.05
75-79	0.27	0.18-0.41	<0.05
80 or older	0.16	0.11-0.24	<0.05
Education level			
Graduated high school	0.9	0.80-1.02	0.11
Attended college or technical school	1.06	0.94-1.20	0.35
Graduated from college or technical school	1.13	0.99-1.29	0.08
Income category			
$15,000 to <$25,000	0.73	0.64-0.84	<0.05
$25,000 to <$35,000	0.65	0.57-0.75	<0.05
$35,000 to <$50,000	0.6	0.53-0.68	<0.05
$50,000 to <$100,000	0.57	0.49-0.66	<0.05
$100,000 to <$200,000	0.46	0.38-0.55	<0.05
$200,000 or more	0.4	0.30-0.54	<0.05

CI: confidence interval, BMI: body mass index, $: US dollar

odds (OR=1.5, 95% CI: [1.37-1.65], p<0.05). Alcohol consumption was not significantly associated with PSD. Diabetes (OR=1.22, 95% CI: [1.13-1.31], p<0.05) and high cholesterol (OR=1.4, 95% CI: [1.30-1.51], p<0.05) were significantly associated with increased odds of PSD in stroke patients. Hypertension was not significantly associated with PSD.

Age group (30.11%) and gender (18.0%) were the most informative features for predicting PSD. Marital status, income category, and physical activity showed moderate MI scores (9-10%), while smoking status, alcohol status, education level, and BMI category had MI scores ranging from 4.74-5.70%. Diabetes status, hypertension status, and cholesterol status had very low MI scores (Figure 1)

Age group (22.64%) and gender (22.41%) had the highest AMI scores, confirming their importance in predicting PSD. Marital status (12.01%), income category (11.50%), and physical activity (11.16%) also showed relatively high AMI values. Smoking status, alcohol status, education level, and BMI category had low AMI values, while diabetes status, hypertension status, and cholesterol status had very low AMI values (Figure 2).

Random forest demonstrated the best performance among the 4 models (accuracy=0.73, precision=0.71, recall=0.77, F1-score=0.74, AUC-ROC=0.81; Table 3). Hyperparameter optimization using GridSearchCV with 5-fold cross-validation further improved performance (accuracy=0.74, precision=0.72, recall=0.78, F1-

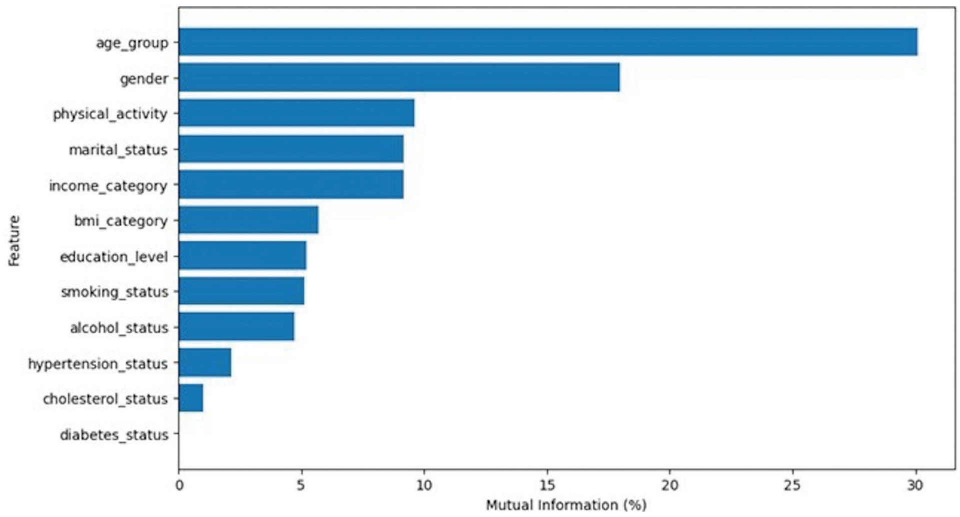

Figure 1 - Mutual information of the features.

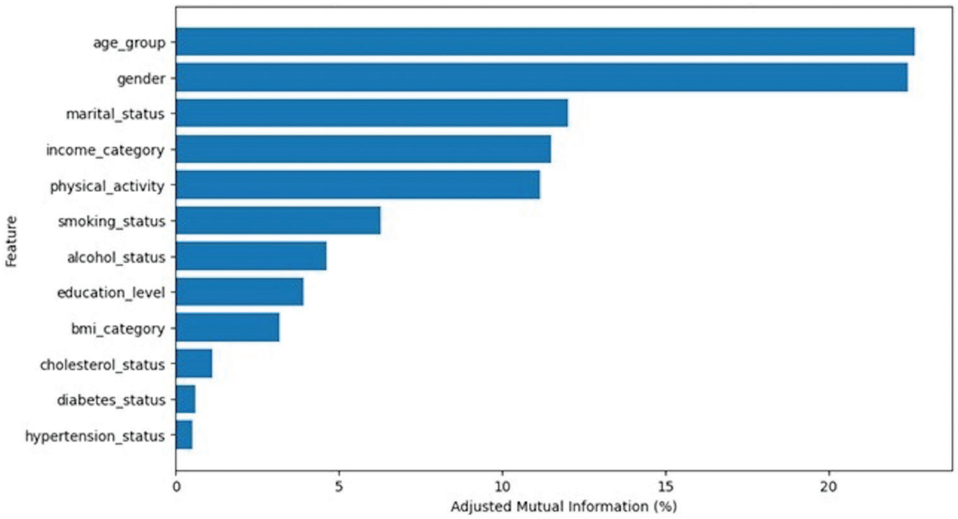

Figure 2 - Adjusted mutual information of the features.

Table 3 - Models evaluation metrics.

Models	Accuracy	Precision	Recall	F1-score	AUC-ROC
Random forest	0.73	0.71	0.77	0.74	0.81
Decision tree	0.70	0.69	0.73	0.71	0.73
Gradient boosting	0.65	0.66	0.64	0.65	0.72
Logistic regression	0.65	0.65	0.62	0.64	0.70

F1-score: harmonic mean of precision and recall,
AUC: area under the curve, ROC: receiver-operating characteristic curve

score=0.75, AUC-ROC=0.81). Precision quantifies the proportion of true positives among all predicted positives, with values above 0.7 generally considered acceptable. Recall indicates the proportion of true positives identified out of all actual positive cases; a score greater than 0.7 is typically desired. The F1-score offers a balanced assessment of model accuracy, where a value exceeding 0.7 suggests good performance. The AUC-ROC measures the model's discriminatory power across different thresholds, while values above 0.7 are acceptable. Therefore, the optimized model's performance was good.

Discussion.

Post-stroke depression is a prevalent complication that significantly hampers the recovery and quality of life of stroke survivors, affecting approximately one-third of this population. The current study aimed to elucidate the multifactorial nature of PSD by analyzing a comprehensive dataset from BRFSS. The findings underscore the complex interplay of sociodemographic, lifestyle, and clinical factors that contribute to the incidence of depression among stroke patients. We employed 2 complementary approaches to assess the relationship between potential risk factors and PSD: logistic regression and AMI. Logistic regression ORs reflect the independent effect of each predictor while controlling all other variables in the model. In contrast, AMI indicates the overall or univariate-level association between a single predictor and the target variable (depression stroke) without necessarily adjusting for the presence of other covariates. Interpreting both together provides a deeper understanding, where variables with high AMI may be strong univariate predictors, whereas variables that are significant in logistic regression indicate robust independent associations once confounders are accounted for.

Sociodemographic factors such as gender, age, marital status, and income level significantly influence depression in stroke survivors. The study identifies age (AMI=22.64%) as the most significant predictor of depression among stroke patients, indicating different depression experiences based on age. Further, we found that increasing age is linked to lower odds of PSD, especially in individuals aged 80 and older. The high AMI confirms the large overall impact of age on depression risk; the logistic regression reinforces that age retains its effect even when controlling for other demographic and clinical variables. This trend is supported by findings that suggest individuals under 75 are more prone to PSD, while those 80 and older have lower chances.[16] This suggests that age may play a protective role against the development of PSD, potentially due to various factors such as increased resilience or differences in coping mechanisms among older adults.[17] Gender follows a similar pattern: the logistic model indicates that males have 43% lower odds of depression, and it also has a high AMI (22.41%), suggesting that this factor is both independently and univariately predictive, which supports previous research indicating a higher depression risk in women after a stroke, regardless of age.[18,19] Nevertheless, other research has indicated an absence of noteworthy gender-based disparities in depression outcomes, implying that although females might endure an elevated incidence of mortality and stroke recurrence, the prevalence rates of depression appear comparable across genders.[20] Marital status (AMI=12.01%) emerges as another protective factor, where married individuals having 14% lower odds of PSD. The protective effect of marital status against PSD, evident in both methods, underscores the importance of social support and relationship status as a buffer against PSD.[37,38] Regarding income, our findings revealed a clear gradient effect, with higher income levels associated with decreased odds of depression. The highest income category showed a 60% reduction in odds compared to the lowest. The AMI of 11.50% for income places it among the top univariate predictors. Education level, despite acceptable univariate correlation (AMI=3.94%), does not remain significant in the multivariable model, suggesting that its effect is largely mediated by other socioeconomic variables such as income. This aligns with the link between socioeconomic status and mental health affecting psychological well-being. For instance, one study found that higher education levels, which often correlate with higher income, were associated with a reduced risk of PSD, particularly in rural areas.[22] Similarly, another study highlighted that low income and primary school education were linked to poorer outcomes in terms of mood and daily living activities post-stroke, suggesting that higher income could mitigate these adverse effects.[23] However, the influence of education on PSD is not uniform across all demographics. Some research suggests that the impact of education on PSD is more pronounced in younger adults compared to older adults, indicating that age may modulate this relationship.[24]

Lifestyle factors were also significant contributors to PSD. Our study found that obesity shows a modest overall association in AMI (3.19%), however, linked to a 33% increased risk of depression, which come with agreement with another study.[25] However, a study reported that underweight was significantly associated with an increased risk of PSD.[20] Further, it has been reported that the impact on PSD varied with age, showing a "U"-shaped curve in younger patients, suggesting that both low and high BMI could be associated with PSD in different age groups.[26] Physical activity exhibits a moderate but consistent protective relationship (OR=0.77, p<0.05) and a relatively high AMI (11.16%), underscoring the potential mental health benefits of remaining active post-stroke. Several studies highlight the positive impact of physical activity on reducing depressive symptoms in stroke survivors.[27,28] Collectively, these findings underscore the importance of incorporating physical activity into stroke rehabilitation programs to mitigate the risk of PSD, improve mental health outcomes, and enhance overall recovery.[29] However, barriers such as fatigue and psychological impairments may hinder exercise participation, necessitating tailored interventions to encourage physical activity among stroke survivors with depression. Smoking, on the other hand, increases the odds of depression (OR=1.50, $p<0.05$) and has a moderate AMI (6.28%), indicating that while it is less influential than age or gender in a univariate sense, it still exerts a clear independent risk in the adjusted model. Smoking has been consistently identified as a significant risk factor for PSD across multiple studies.30-32 Overall, the evidence underscores the importance of smoking cessation interventions as part of post-stroke care to mitigate the risk of depression and improve overall outcomes for stroke survivors. Alcohol use displays a minor univariate association (AMI=4.62%) but no significant effect when other variables are considered (OR=0.97, $p=0.43$), implying that its impact on depression is subsumed by related lifestyle or demographic factors. The relationship between alcohol consumption and stroke risk appears to be J-shaped, with light-to-moderate drinking associated with decreased risk of ischemic stroke, while heavy drinking increases the risk.[33]

Among comorbidities, diabetes exemplifies the difference between univariate and multivariable assessments, showing only a small AMI (0.60%) yet an elevated OR in the logistic model (OR=1.22, $p<0.05$). This gap suggests that diabetes becomes more salient once other factors are held constant. Cholesterol also has a relatively low AMI (1.13%) but remains independently associated with depression (OR=1.40, $p<0.05$), indicating that elevated cholesterol may subtly increase depression risk in stroke patients beyond what is captured by simpler univariate measures. Finally, hypertension does not appear to be related to depression in either approach (OR=0.96, $p=0.30$; AMI=0.52%), pointing to its minimal role compared to the other comorbid conditions. Overall, the contrast between logistic regression and AMI highlights how certain factors, such as diabetes or cholesterol, may display modest univariate effects yet have important independent relationships when demographic, socioeconomic, and lifestyle confounders are addressed. The relationship between diabetes and PSD appears to be complex and somewhat contradictory based on the studies provided. It has been found that the presence of diabetes was significantly associated with PSD in their multivariable regression analysis.[34] This supports diabetes as a potential predictor. In contrast, other studies reported that having a comorbid diagnosis of diabetes was associated with lower odds of having lower depression severity.[35] This implies diabetes may actually be protective against PSD. The inconsistency in findings across studies suggests that the relationship between diabetes and PSD may be influenced by other factors or may vary depending on the specific patient population and study methodology. Furthermore, a history of hyperlipidemia has been identified as a predictive factor for PSD, emphasizing the importance of monitoring and managing cholesterol levels in stroke patients.[36] While pre-existing hypertension was not identified as a significant predictor, new-onset hypertension following stroke was found to be an important feature in PSD prediction models.[39]

Among the evaluated models, random forest demonstrated the highest overall performance, achieving an accuracy of 0.73 and an AUC-ROC of 0.81 in predicting PSD. Hyperparameter optimization enhanced model performance, and cross-validation on the BRFSS dataset confirmed this improvement and its internal consistency. Future studies could explore the integration of additional data sources, such as neuroimaging or genetic data, to further improve the accuracy of predictive models.

A random forest model demonstrated strong performance in predicting PSD, achieving high accuracy (0.73) and AUC-ROC (0.81) using sociodemographic, lifestyle, and clinical factors. This result aligns with the growing body of literature supporting the potential of machine learning in PSD prediction. This performance compares favorably with other machine learning approaches for PSD prediction. For instance, Chen et al[39] achieved specificities of 0.83-0.91 and sensitivities of 0.30-0.48 using clinical features from a large cohort of ischemic stroke patients. Other studies have explored different data modalities, such as serum biomarkers like interleukin.[10,40] While these varied approaches demonstrate the potential of machine learning for PSD prediction, this study's random forest model, with its robust performance and potential for enhancement through integration of neuroimaging or genetic data, offers a promising avenue for early identification and intervention in PSD.

Study limitations. First, the data is cross-sectional, which limits our ability to draw causal inferences. Second, the BRFSS relies on self-reported data, which may be subject to recall bias and social desirability bias. Third, the data is limited to variables collected in the BRFSS and may not include all potential confounders. Finally, the generalizability of the findings may be limited by the specific characteristics of the BRFSS sample. However, it's important to note that broader generalization can be achieved through external validation using datasets outside the scope of this paper. Future studies employing diverse populations and independent datasets will be essential to confirm the robustness and applicability of our model. In addition, future research should consider longitudinal studies with more comprehensive data collection to further investigate the factors associated with PSD.

In conclusion, the findings of this study illuminate the multifactorial nature of PSD, emphasizing the need for a comprehensive approach to assessment and intervention that addresses both individual and contextual factors. Early identification and effective management of PSD are crucial in enhancing recovery and quality of life for stroke survivors.

Acknowledgment. *The author gratefully acknowledge Cambridge Proofreading LLC (www.cambridgeproofreading.com) for the English language editing.*

References

1. Aslan IK, Akpinar A, Salt I. Evaluation of depression frequency and its effect on prognosis in patients treated for acute ischemic stroke. *MNJ Malang Neurol J* 2024; 10: 101-107.
2. Dai J, Zhao SS, Zhang SX. Early screening for post-stroke depression and its effect on functional outcomes, quality of life, and mortality: a meta-analysis. *World J Psychiatry* 2024; 14: 1397-1403.
3. Fischer S, Linseisen J, Kirchberger I, Zickler P, Ertl M, Naumann M, et al. Association of post-stroke-depression and health-related quality of life 3 months after the stroke event. Results from the Stroke Cohort Augsburg (SCHANA) study. *Psychol Health Med* 2023; 28: 1148-1159.
4. Liu L, Li X, Marshall IJ, Bhalla A, Wang Y, O'Connell MDL. Trajectories of depressive symptoms 10 years after stroke and associated risk factors: a prospective cohort study. *Lancet* 2023; 402: S64.
5. Ladwig S, Werheid K, Südmeyer M, Volz M. Predictors of post-stroke depression: validation of established risk factors and introduction of a dynamic perspective in 2 longitudinal studies. *Front Psychiatry* 2023; 14: 1093918.
6. Almeida OP, Hankey GJ, Ford AH, Etherton-Beer C, Flicker L, Hackett ML. Measures associated with early, late, and persistent clinically significant symptoms of depression one year after stroke in the AFFINITY trial. *Neurology* 2022; 98: e1021-e1030.
7. Senff J, Tack RWP, Mallick A, Gutierrez-Martinez L, Duskin J, Kimball TN, et al. Modifiable risk factors for stroke, dementia and late-life depression: a systematic review and DALY-weighted risk factors for a composite outcome. *J Neurol Neurosurg Psychiatry* 2025: jnnp-2024-334925.
8. Kunugi H. Depression and lifestyle: focusing on nutrition, exercise, and their possible relevance to molecular mechanisms. *Psychiatry Clin Neurosci* 2023; 77: 420-433.
9. Govori V, Budinčević H, Morović S, Đerke F, Demarin V. Updated perspectives on lifestyle interventions as secondary stroke prevention measures: a narrative review. *Medicina (Kaunas)* 2024; 60: 504.
10. Tikk K, Sookthai D, Monni S, Gross ML, Lichy C, Kloss M, et al. Primary preventive potential for stroke by avoidance of major lifestyle risk factors: the European Prospective Investigation into Cancer and Nutrition-Heidelberg cohort. *Stroke* 2014; 45: 2041-2046.
11. Dulay MF, Criswell A, Hodics TM. Biological, psychiatric, psychosocial, and cognitive factors of poststroke depression. *Int J Environ Res Public Health* 2023; 20: 5328.
12. Butt TI, Shazdi K, Anjum F, Yaqoob MA, Sattar MMM, Umer A. Review the current therapeutic interventions and management strategies that target both mental health and metabolic abnormalities in individuals. *J Popul Ther Clin Pharmacol* 2024; 31: 418-422.
13. Del Zotto E, Costa P, Morotti A, Poli L, de Giuli V, Giossi A, et al. Stroke and depression: a bidirectional link. *World J Meta-Anal* 2014; 2: 49-63.
14. Hoertel N, Limosin F. Poststroke depression and major depressive disorder: the same or separate disorders? *Int Psychogeriatr* 2020; 32: 1279-1281.
15. Trusova NA, Levin OS. [Clinical significance and possibilities of therapy of post-stroke depression]. *Zh Nevrol Psikhiatr Im S S Korsakova* 2019; 119: 60-67. [In Russian].
16. Mayman N, Stein LK, Erdman J, Kornspun A, Tuhrim S, Jette N, et al. Risk and predictors of depression following acute ischemic stroke in the elderly. *Neurology* 2021; 96: e2184-e2191.
17. Zhou X, Liu Z, Zhang W, Zhou L. Resilience is associated with post-stoke depression in Chinese stroke survivors: a longitudinal study. *J Affect Disord* 2020; 273: 402-409.
18. Dymm B, Goldstein LB, Unnithan S, Al-Khalidi HR, Koltai D, Bushnell C, et al. Depression following small vessel stroke is common and more prevalent in women. *J Stroke Cerebrovasc Dis* 2024; 33: 107646.
19. Kim M, Lee YH. Gender differences in the risk of depression in community-dwelling stroke survivors compared to the general population without stroke. *Chonnam Med J* 2023; 59: 134-139.
20. Guo X, Xiong Y, Huang X, Pan Z, Kang X, Chen C, et al. Gender-based differences in long-term outcomes after stroke: a meta-analysis. *PLoS One* 2023; 18: e0283204.
21. Wang B, Ding XX, Zhang H, Liu ZM, Duan PB, Dong YF. Predictors of post-stroke depression: the perspective from the social convoy model. *Psychogeriatrics* 2023; 23: 864-875.
22. Cai Q, Qian M, Chen M. Association between socioeconomic status and post-stroke depression in middle-aged and older adults: results from the China health and retirement longitudinal study. *BMC Public Health* 2024; 24: 1007.
23. Lindmark A, von Euler M, Glader EL, Sunnerhagen KS, Eriksson M. Socioeconomic differences in patient reported outcome measures 3 months after stroke: a nationwide swedish register-based study. *Stroke* 2024; 55: 2055-2065.
24. Samudio-Cruz MA, Toussaint-González P, Estrada-Cortés B, Martínez-Cortéz JA, Rodríguez-Barragán MA, Hernández-Arenas C, et al. Education level modulates the presence of poststroke depression and anxiety, but it depends on age. *J Nerv Ment Dis* 2023; 211: 585-591.
25. Meshkat S, Tassone VK, Wu M, Duffy SF, Boparai JK, Jung H, et al. Does self-reported BMI modify the association between stroke and depressive symptoms? *Can J Neurol Sci* 2025; 52: 68-74.
26. Xue Z, Wang Y, Wang L, Shen L, Zhang A, Pan P, et al. Analysis of influencing factors of poststroke depression: is higher body mass index always a risk factor of poststroke depression? *J Nerv Ment Dis* 2019; 207: 203-208.
27. Apriliyasari RW, Budi IS, Tan MP, Tsai PS. Physical activity and depression in Indonesian adults with stroke: a nationwide survey. *J Nurs Scholarsh* 2023; 55: 356-364.

28. Chen R, Guo Y, Kuang Y, Zhang Q. Effects of home-based exercise interventions on post-stroke depression: a systematic review and network meta-analysis. ***Int J Nurs Stud*** 2024; 152: 104698.
29. Fauzi LA, Kushartanti W, Arovah NI, Fauzi, Maria R, Anwar A. Investigating the relationship between physical activity and depression level with stroke recurrences: an observational cross-sectional study. ***Fizjoterapia Pol*** 2024; 24: 210-215.
30. Khedr EM, Abdelrahman AA, Desoky T, Zaki AF, Gamea A. Post-stroke depression: frequency, risk factors, and impact on quality of life among 103 stroke patients - hospital-based study. ***Egypt J Neurol Psychiatry Neurosurg*** 2020; 56: 66.
31. Parikh NS, Salehi Omran S, Kamel H, Elkind MSV, Willey J. Symptoms of depression and active smoking among survivors of stroke and myocardial infarction: an NHANES analysis. ***Prev Med*** 2020; 137: 106131.
32. Rabat Y, Sibon I, Berthoz S. Implication of problematic substance use in poststroke depression: an hospital-based study. ***Sci Rep*** 2021; 11: 13324.
33. Jeong SM, Lee HR, Han K, Jeon KH, Kim D, Yoo JE, et al. Association of change in alcohol consumption with risk of ischemic stroke. ***Stroke*** 2022; 53: 2488-2496.
34. Krinock MJ, Singhal NS. Diabetes, stroke, and neuroresilience: looking beyond hyperglycemia. ***Ann N Y Acad Sci*** 2021; 1495: 78-98.
35. Waller A, Fakes K, Carey M, Dizon J, Parrey K, Coad M, et al. Quality of life and mood disorders of mild to moderate stroke survivors in the early post-hospital discharge phase: a cross-sectional survey study. ***BMC Psychol*** 2023; 11: 32.
36. Xiong B, Li Z, Zhang S, Wang Z, Xie Y, Zhang M, et al. Association between non-high-density lipoprotein cholesterol to high-density lipoprotein cholesterol ratio (NHHR) and the risk of post-stroke depression: a cross-sectional study. ***J Stroke Cerebrovasc Dis*** 2024; 33: 107991.
37. Wang Y, Zha F, Han Y, Cai Y, Chen M, Yang C, et al. Nonlinear connection between remnant cholesterol and stroke risk: evidence from the China health and retirement longitudinal study. ***Lipids Health Dis*** 2023; 22: 181.
38. Yuan Y, Zhou X, Jia W, Zhou J, Zhang F, Du J, et al. The association between self-monitoring of blood glucose and HbA1c in type 2 diabetes. ***Front Endocrinol (Lausanne)*** 2023; 14: 1056828.
39. Chen YM, Chen PC, Lin WC, Hung KC, Chen YB, Hung CF, et al. Predicting new-onset post-stroke depression from real-world data using machine learning algorithm. ***Front Psychiatry*** 2023; 14: 1195586.
40. Chi CH, Huang YY, Ye SZ, Shao MM, Jiang MX, Yang MY, et al. Interleukin-10 level is associated with post-stroke depression in acute ischaemic stroke patients. ***J Affect Disord*** 2021; 293: 254-260.