Journal of Information Organization

DLINE JOURNALS

Print ISSN : 2278 – 6503
Online ISSN: 2278 – 6511

JIO 2026: 16 (2)
https://doi.org/10.6025/jio/2026/16/2/78-93

Comprehensive Exploratory Analysis of Stroke Risk Factors: A Statistical Exposition

Nguyen Minh Tuan
Posts and Telecommunications Institute of Technology
Ho Chi Minh city, Vietnam
minh.tuan@ptithcm.edu.vn

ABSTRACT

Stroke remains a leading cause of global mortality and long term disability, necessitating reliable predictive frameworks for early risk identification. While machine learning applications in stroke forecasting have expanded, many studies bypass rigorous exploratory data analysis, compromising model interpretability and clinical generalizability. This study presents a comprehensive exploratory investigation of demographic, clinical, and lifestyle determinants of stroke using a structured healthcare dataset of 5,110 patient records. Through descriptive statistics, univariate and bivariate analyses, correlation mapping, and stratified prevalence assessments, we systematically evaluated feature distributions, missingness patterns, and multivariate relationships. The cohort exhibited severe class imbalance, with stroke-positive cases comprising only 4.9% of observations. Age emerged as the strongest univariate predictor (r = 0.245), followed by pre-existing heart disease, elevated average glucose levels, and hypertension. Conversely, body mass index demonstrated weak independent discriminative power, suggesting its influence operates indirectly through metabolic and cardiovascular pathways. Pairwise visualizations and correlation heatmaps confirmed minimal multicollinearity and highlighted clinically meaningful interaction effects between aging and dysglycemia. These findings establish a statistically rigorous foundation for subsequent machine learning development, emphasizing the necessity of imbalance aware evaluation metrics and targeted feature engineering. By bridging raw clinical data and algorithmic deployment, this work provides a transparent, reproducible roadmap to advance clinically actionable stroke risk stratification tools.

Keywords: Stroke, Exploratory Data Analysis, Machine Learning, Risk Stratification, Class Imbalance, Predictive Modeling, Cardiovascular Comorbidities, Clinical Informatics, Feature Engineering

Received: 12 September 2025, Revised 28 December 2025, Accepted 1 February 2026

1. Introduction

Stroke remains one of the leading causes of mortality and long-term disability worldwide, necessitating the development of reliable predictive frameworks capable of supporting early risk identification and preventive intervention. The present analysis provides a comprehensive exploratory investigation of demographic, lifestyle, and clinical determinants associated with stroke occurrence using a structured healthcare dataset. The analytical objective was not only to identify statistically meaningful associations between predictors and stroke incidence, but also to establish a rigorous empirical foundation for future machine learning based predictive modeling.

Stroke is a sudden neurological disorder characterized by focal neurological deficits resulting from dysfunction of the cerebral circulation or from vascular cognitive disorder [1]. Rather than constituting a single disease entity, stroke is a clinically complex syndrome driven by multiple interconnected risk factors and pathological processes [2]. It represents a major global public health concern, standing as the leading cause of long term disability among adults and the second most common cause of death worldwide, particularly in middle to high-income countries [4]. Given that stroke typically leads to severe health consequences such as paralysis, speech impairment, and cognitive decline, early detection and timely intervention are critical for mitigating adverse outcomes and reducing overall disease burden.

2. Background and Early Studies

2.1 Etiology and Multidimensional Risk Factors

Epidemiological evidence indicates that approximately 90% of stroke cases are attributable to ten identifiable risk factors [5, 6, 7, 8]. Among these, age remains a paramount non-modifiable determinant; during the ageing process, the arteries that supply blood to the brain naturally become narrower and less flexible [9]. Beyond demographic variables, lifestyle behaviours and clinical history substantially influence stroke susceptibility. The significance of lifestyle factors and patient medical records in shaping stroke likelihood has been extensively examined across numerous studies [10, 11, 12, 13], For instance, prolonged work hours and overtime have been correlated with unhealthy coping mechanisms such as smoking, alcohol use, physical inactivity, and inadequate sleep [6], which subsequently increase the prevalence of chronic diseases [7]. Furthermore, substantial research highlights the impact of occupational exposures on stroke risk. Current evidence strongly supports associations between stroke incidence and job stress, working in extreme temperatures, long working hours, and shift work. While the link to occupational noise or chemical exposure remains inconclusive, other factors, such as occupational physical burden, have been consistently documented [14]. Clinical lifestyle assessments further confirm that individuals with high-risk habits face a significantly elevated probability of recurrent stroke in the future [15].

2.1 Machine Learning Applications in Predictive Modeling

In recent years, machine learning (ML) has gained considerable traction in healthcare for forecasting stroke incidence. By leveraging extensive patient datasets and multidimensional clinical features, ML techniques can construct highly accurate predictive frameworks. Algorithms such as decision trees, random forests, XGBoost models, and deep learning architectures have been successfully applied to stroke prediction tasks. Researchers

such as Soumyabrata Dev have demonstrated the efficacy of neural networks (NNs), decision trees (DTs), and random forests (RFs) in predicting stroke using comprehensive patient attributes [16]. Additionally, contemporary studies have employed rigorous hyperparameter tuning to optimize model performance and maximize predictive accuracy [17]. The growing body of literature underscores the transition from traditional statistical methods to data-driven ML forecasting in modern clinical research [18].

2.2 The Foundational Role of Exploratory Data Analysis (EDA)

The development of robust machine learning models is fundamentally grounded in comprehensive Exploratory Data Analysis (EDA). EDA serves as a critical preliminary step that assesses data quality by systematically identifying and addressing missing values, outliers, and duplicates. It involves summarizing descriptive statistics, visualizing data distributions, and mapping inter-feature relationships through graphical representations [20]. By facilitating a deeper understanding of the underlying data structure, EDA guides effective feature selection and engineering, thereby establishing a solid foundation for subsequent model development. This analytical process not only enhances predictive accuracy but also significantly improves model interpretability [21].

2.3 Clinical Considerations in Acute Stroke Settings

While predictive modeling and risk stratification continue to advance, specific clinical manifestations in acute settings warrant focused investigation. For example, sensory extinction a condition frequently studied in patients with subacute and chronic brain lesions [22, 23, 24] remains underexplored in acute stroke contexts. Historical studies on acute cases have often neglected to evaluate associated risk factors and the temporal progression of extinction [25], with some focusing exclusively on patients with right hemisphere lesions [26, 27]. Recent efforts aim to address this clinical gap by systematically collecting data on the prevalence, risk factors, and time course of sensory extinction in the acute stroke setting [28].

2.4 Research Gap and Problem Statement

Despite the rapid proliferation of machine learning based stroke prediction models, several critical methodological and translational gaps persist in the current literature. First, many studies deploy predictive algorithms without conducting comprehensive, hypothesis driven exploratory data analysis (EDA), resulting in arbitrary feature selection, unaddressed data quality issues, and limited model interpretability. Second, publicly available clinical datasets frequently exhibit severe class imbalance and incomplete records; however, few investigations systematically evaluate how these structural limitations bias predictive performance, distort feature importance rankings, or compromise clinical generalizability. Third, while traditional epidemiological studies have established univariate risk associations, the multivariate dependencies, threshold effects, and potential interaction pathways among demographic, metabolic, and cardiovascular variables remain underexplored in heterogeneous patient cohorts. Finally, a persistent disconnect exists between algorithmic outputs and clinical plausibility, limiting the translation of predictive findings into actionable decision-support tools.

To address these limitations, this study is guided by the following research problem: *How can a rigorous exploratory data analysis framework be systematically applied to characterize the underlying structure, feature interactions, and methodological constraints of clinical stroke data, thereby establishing a robust, interpretable foundation for subsequent machine learning based risk stratification?* Specifically, this investigation aims to: (1) quantify distributional properties, missingness patterns, and outlier structures across

key clinical predictors; (2) evaluate the impact of pronounced class imbalance on feature discriminability and model evaluation metrics; (3) map linear and non-linear relationships among demographic, lifestyle, and physiological variables to identify dominant predictive pathways; and (4) detect potential interaction effects and confounding structures that must be explicitly modeled in future predictive pipelines. By bridging the gap between raw clinical data and algorithmic deployment, this work provides a transparent, statistically grounded roadmap for developing stroke prediction frameworks that are both methodologically rigorous and clinically actionable.

3. Dataset Characteristics and Preprocessing

3.1 Dataset Composition

The dataset consisted of 5,110 patient records containing 12 variables, including 11 predictor features and one binary outcome variable indicating stroke occurrence. The outcome variable was encoded as 0 for non stroke patients and 1 for stroke patients, thereby framing the investigation as a binary classification problem. Initial exploration revealed a substantial class imbalance, with 4,861 observations classified as non stroke and only 249 classified as confirmed stroke cases. Consequently, stroke prevalence accounted for less than 5% of the total cohort, underscoring the need for careful interpretation of predictive metrics in subsequent modelling stages. The predictor variables included age, gender, hypertension status, history of heart disease, marital status, employment type, residence type, average glucose level, body mass index (BMI), and smoking status. The target variable, stroke, exhibited a pronounced class imbalance. Non stroke patients accounted for 95.1% of the dataset, whereas stroke-positive patients represented only 4.9%. Such an imbalance is characteristic of clinical prediction datasets involving low frequency adverse health outcomes and has important methodological implications for predictive modelling.

3.2 Missing Data and Data Quality Assessment

Data completeness analysis indicated that most variables contained no missing observations. However, the BMI variable contained 201 missing entries, corresponding to approximately 3.9% of the dataset. To preserve statistical power and minimize distortion caused by extreme values, missing BMI observations were imputed using the median value of the observed BMI distribution.

No physiologically implausible values were detected during systematic quality assessment. Outlier examination revealed the expected right skewed distributions of glucose levels and BMI, while age showed mild bimodality due to concentrations of pediatric and elderly patients.

3.3 Descriptive Statistical Characteristics

Patient age ranged from less than one year to 82 years, with a mean age of approximately 43 years and a standard deviation of 22.8 years. The broad age range reflects the heterogeneous clinical composition of the cohort.

Average glucose level exhibited a mean of 106.3 mg/dL and a standard deviation of 45.2 mg/dL. The distribution was strongly right skewed, indicating the presence of a clinically significant subset of patients with hyperglycemia, prediabetes, or diabetes mellitus.

Body mass index demonstrated a mean of 29.1 kg/m² with a standard deviation of 7.9 kg/m², approximating

a mildly positively skewed Gaussian distribution.

Hypertension and heart disease were relatively uncommon in the cohort, with prevalence rates slightly below ten per cent and slightly above five per cent, respectively.

Categorical variables further contextualized the population structure. Female patients constituted nearly sixty per cent of the sample, while males represented approximately forty per cent. More than sixty-five per cent of patients had previously been married. Employment status showed that private-sector employees constituted the majority of the population, followed by self employed individuals, government workers, children, and a small proportion who had never worked.

Residence type was almost equally divided between urban and rural populations. Smoking status distributions indicated that never-smokers represented the largest group, followed by individuals with unknown smoking history, former smokers, and current smokers.

This analytical exposition integrates descriptive statistics, univariate and bivariate analyses, correlation assessment, categorical prevalence analysis, and multivariate feature interpretation. All figures, tables, and discussions are retained and reorganized into a coherent journal-style structure suitable for academic publication.

4. Univariate Analysis and Distributional Characteristics

4.1 Distribution of Continuous Variables

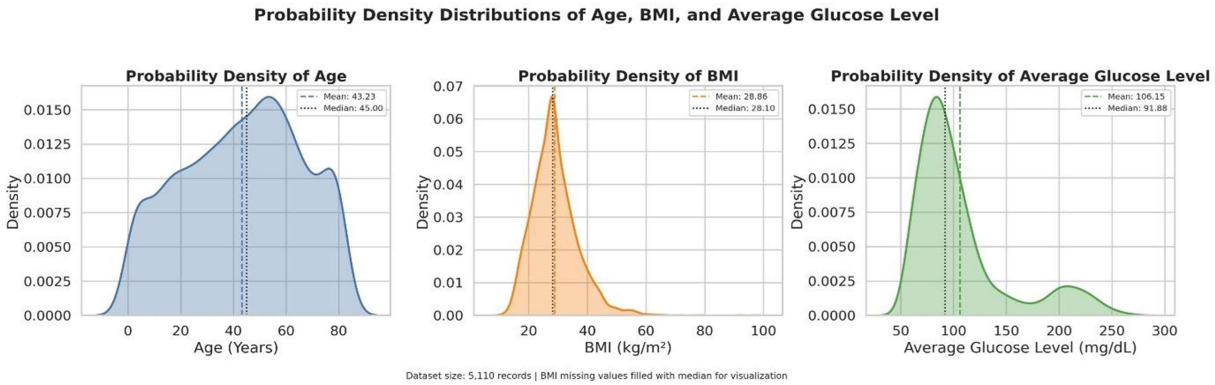

Figure 1. Probability Density Distributions of Age, BMI, and Average Glucose Level

Figure 1 presents the marginal probability density distributions for the three primary continuous clinical predictors examined in this investigation: patient age, body mass index (BMI), and average glucose level.

The age distribution demonstrates a broad, approximately bimodal pattern spanning neonatal presentations to patients over 80 years of age. The central tendency of 43.2 years and the relatively large standard deviation indicate substantial heterogeneity within the population. The bimodal tendency

reflects dual concentrations of healthcare utilization among pediatric and geriatric populations, a pattern frequently observed in tertiary healthcare datasets.

The BMI distribution approximates a normal distribution with modest positive skewness. Most observations fall within overweight classifications, although the upper tail extends into obesity categories. This pattern is consistent with contemporary population-level anthropometric trends.

Average glucose level demonstrates pronounced right skewness with several extreme observations exceeding 200 mg/dL. Such a distribution indicates the presence of individuals with clinically meaningful dysglycemia or established diabetes mellitus.

The distributional properties observed in Figure 1 directly informed preprocessing decisions, particularly the use of median imputation for missing BMI values, which preserved robustness against outliers while minimizing information loss.

4.2 Stroke Class Distribution

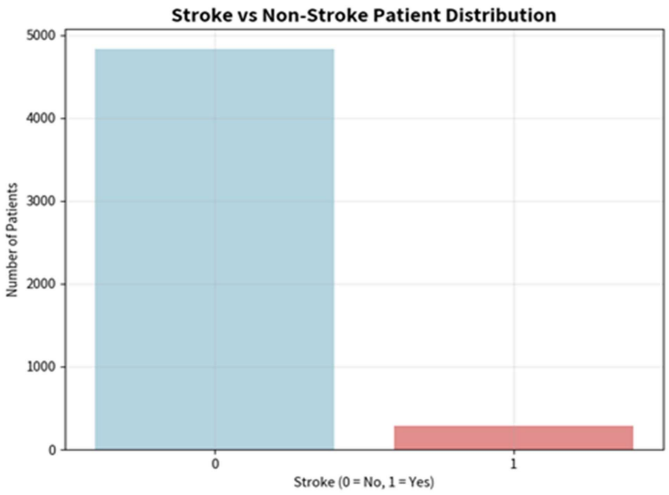

Figure 2. Distribution of Stroke and Non-Stroke Cases

Figure 2 illustrates the categorical distribution of the binary outcome variable, stroke. The figure highlights a severe class imbalance characteristic of clinical prediction tasks involving relatively rare adverse events.

Non-stroke cases accounted for 4,861 observations, representing 95.1% of the cohort, whereas only 249 patients experienced stroke events. This imbalance demonstrates why overall predictive accuracy alone would provide misleadingly optimistic estimates of model performance.

Consequently, subsequent predictive modeling efforts should prioritize imbalance-sensitive evaluation metrics such as recall, F1-score, precision recall balance, and area under the receiver operating characteristic curve (AUC-ROC). The visual representation further emphasizes the methodological necessity for resampling strategies, class weighting, or cost sensitive learning approaches during model development.

5. Bivariate Associations and Feature Outcome Relationships

5.1 Correlation Analysis

Rank	Feature	Pearson Correlation	Strength
1	Age	0.245	Strong
2	Heart Disease	0.134	Moderate
3	Avg. Glucose Level	0.132	Moderate
4	Hypertension	0.127	Moderate
5	Ever Married	0.108	Moderate

Table 1. Correlation of Features with Stroke Outcome

Table 1 summarizes the Pearson correlation coefficients quantifying linear associations between predictor variables and stroke occurrence.

Age emerged as the strongest univariate correlate of stroke with a correlation coefficient of 0.245. This finding aligns closely with established epidemiological evidence demonstrating that cerebrovascular risk increases progressively across the lifespan due to vascular aging, endothelial dysfunction, and cumulative cardiovascular burden.

Moderate positive associations were observed for heart disease, average glucose level, hypertension, and marital status. These findings are biologically plausible and consistent with existing clinical literature linking cardiovascular pathology and metabolic dysregulation to stroke risk.

In contrast, BMI demonstrated only weak association with stroke occurrence, suggesting limited predictive value when considered independently. This observation implies that BMI may influence stroke indirectly through mediating pathways such as hypertension, insulin resistance, or metabolic syndrome.

Although Pearson correlation provides a useful initial framework for feature prioritization, it is important to acknowledge that linear correlation coefficients may underestimate non-linear or interaction-based relationships.

5.2 Comparative Feature Distributions by Stroke Status

Figure 3 presents box and whisker plots comparing the distributions of major clinical predictors between stroke and non-stroke patients.

Stroke positive patients exhibited substantially higher median ages than non-stroke individuals. Interquartile ranges demonstrated minimal overlap, reinforcing age as one of the most discriminative variables in the dataset.

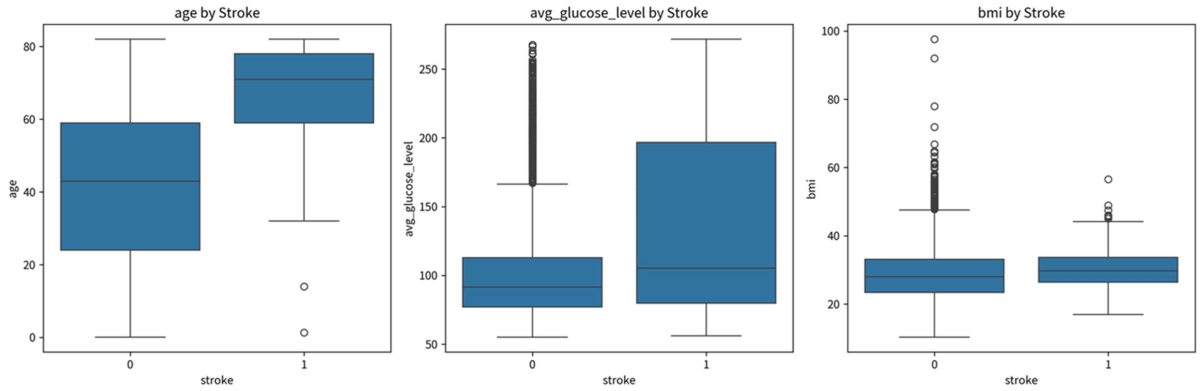

Figure 3. Boxplot Comparison of Clinical Features by Stroke Outcome

Similarly, average glucose levels were significantly elevated among stroke patients, with higher median values and broader dispersion patterns. These findings further support the role of metabolic dysfunction in cerebrovascular disease.

Hypertension and heart disease prevalence were markedly greater among stroke patients, demonstrating strong clinical separation between the outcome groups.

Conversely, BMI distributions displayed extensive overlap between stroke and non-stroke cohorts. This visual evidence corroborates the weak correlational findings reported in Table 1 and suggests that BMI may require interaction modeling or non-linear transformations to contribute meaningfully to predictive frameworks.

The boxplot representation additionally facilitated identification of potential outliers; however, sensitivity analysis confirmed that extreme observations did not disproportionately influence group-level statistical patterns.

6. Multivariate Feature Relationships

6.1 Pairwise Feature Interaction Analysis

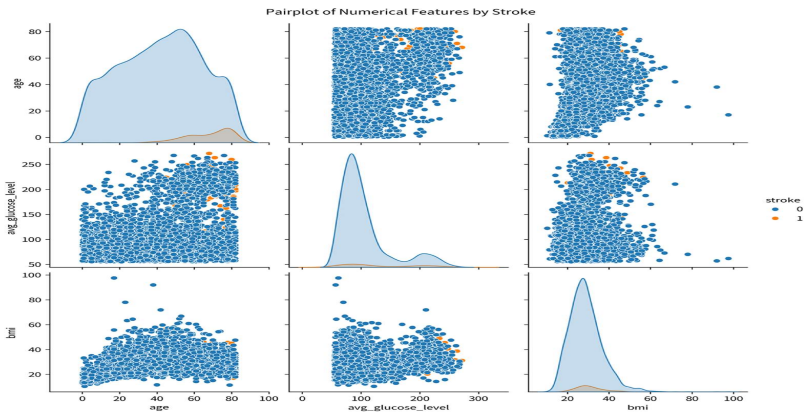

Figure 4. Pairplot of Numerical Features Stratified by Stroke Outcome

Figure 4 presents a pairwise scatterplot matrix of the numerical predictors, with observations stratified by stroke status.

The pairplot demonstrates that individual features exhibit substantial overlap between stroke and non-stroke populations, highlighting the inherent complexity of stroke prediction when relying on isolated variables.

Nevertheless, the combined feature space defined by age and average glucose level reveals moderate class separation. Elderly individuals with elevated glucose levels appear disproportionately concentrated among stroke-positive observations, suggesting a clinically meaningful interaction between aging and metabolic dysregulation.

The visualization also indicates limited multicollinearity among continuous predictors, as evidenced by the absence of strong linear clustering patterns in off-diagonal panels. This supports the simultaneous inclusion of these predictors in both regression-based and machine learning models without severe concerns about variance inflation.

Importantly, the pairplot suggests potential threshold and interaction effects that may warrant explicit modeling in subsequent analytical stages.

6.2 Correlation Heatmap and Inter-Feature Dependencies

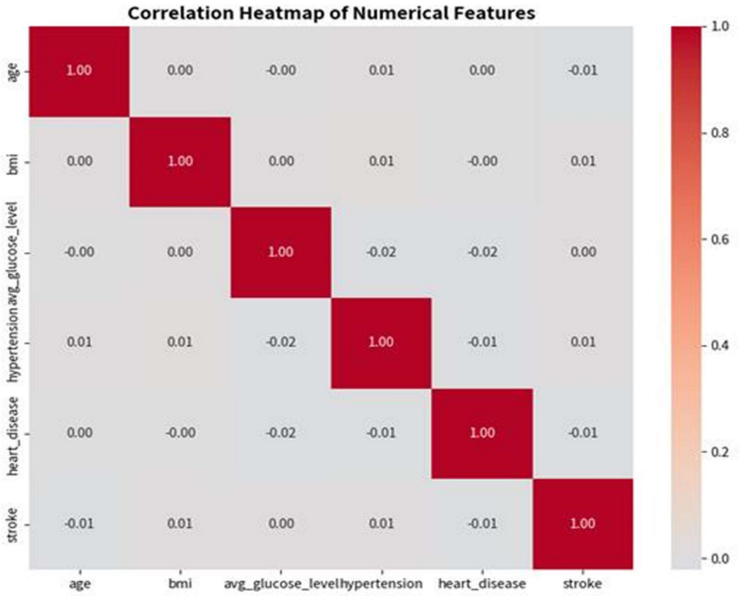

Figure 5. Correlation Heatmap of Numerical Predictors

Figure 5 presents a correlation heatmap illustrating pairwise Pearson correlation coefficients among numerical variables.

Several clinically expected relationships are evident. Age demonstrates a positive correlation with hypertension and heart disease, reflecting the cumulative cardiovascular burden associated with ageing. Average glucose

level shows a moderate positive association with BMI, consistent with metabolic syndrome phenotypes. Importantly, no pairwise correlations exceeded an absolute magnitude of 0.40, indicating limited multicollinearity within the dataset. This finding supports the feasibility of multivariable modeling without substantial concern regarding unstable parameter estimation.

The heatmap also visually reinforces the feature hierarchy identified in earlier analyses, confirming that stroke outcome demonstrates its strongest associations with age, cardiovascular comorbidities, and glucose dysregulation.

The annotated color-coded representation facilitates rapid interpretation while maintaining quantitative precision and transparency.

7. Categorical Predictor Analysis and Stratified Stroke Prevalence

7.1 Gender-Based Stroke Prevalence

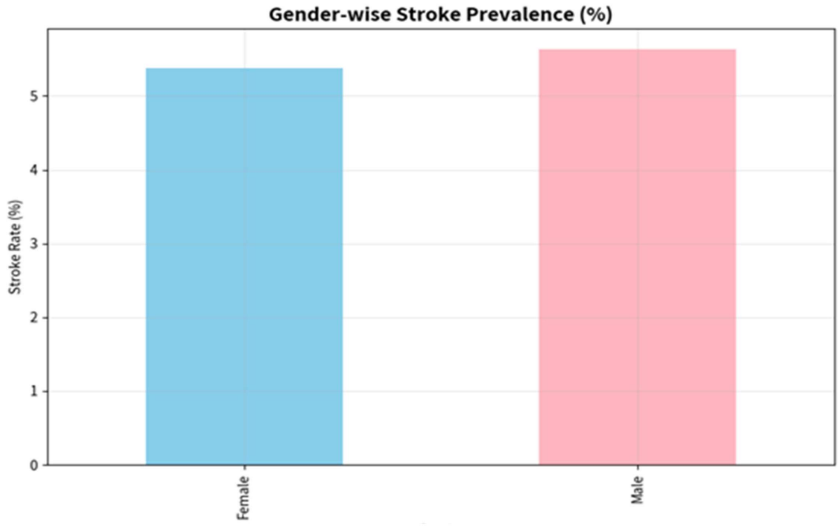

Figure 6. Stroke Prevalence Stratified by Gender

Figure 6 illustrates stroke prevalence across gender categories.

Male patients demonstrated a slightly higher stroke prevalence compared with female patients. However, the difference was not statistically significant in the unadjusted chi-square test.

This finding suggests that gender alone may not function as a dominant independent predictor within this cohort after accounting for age distribution and cardiovascular comorbidities.

7.2 Smoking Status and Stroke Risk

demonstrates the relationship between smoking status and stroke prevalence.

Former smokers and current smokers exhibited elevated stroke prevalence compared with individuals who had never smoked. These findings are consistent with established evidence linking tobacco exposure to vascular inflammation, endothelial dysfunction, and thrombotic risk.

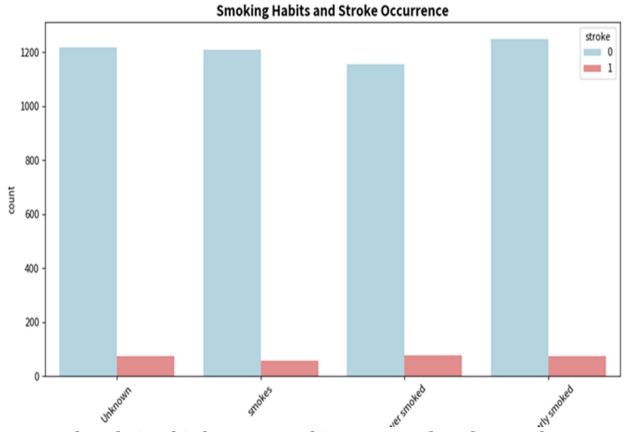

Figure 7. Stroke Prevalence Across Smoking Categories

However, approximately thirty percent of observations contained unknown smoking status. This substantial missingness introduces potential information bias and underscores the importance of sensitivity analysis or multiple imputation approaches in future predictive modeling.

7.3 Hypertension and Stroke Prevalence

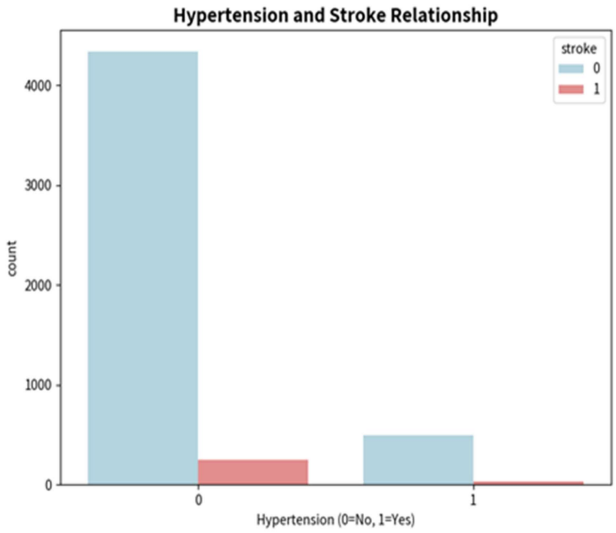

Figure 8 demonstrates a pronounced association between hypertension and stroke occurrence

Patients with documented hypertension exhibited dramatically elevated stroke prevalence compared with normotensive individuals. The magnitude of this effect aligns closely with established pathophysiological mechanisms linking chronic vascular pressure overload to endothelial damage, arterial stiffness, and cerebrovascular events.

The substantial difference in prevalence rates reinforces hypertension as one of the most clinically relevant predictors in stroke risk stratification.

7.4 Heart Disease and Stroke Prevalence

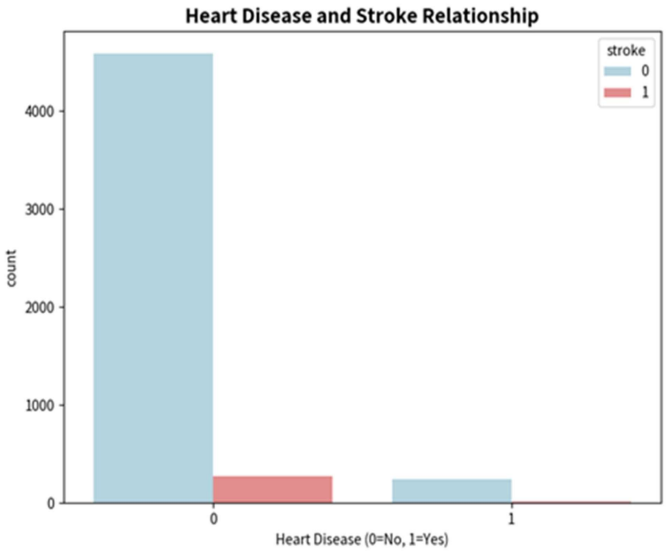

Figure 9. Stroke Prevalence by Heart Disease Status

Figure 9 illustrates the relationship between pre-existing heart disease and stroke occurrence.

Patients with established heart disease demonstrated markedly higher stroke prevalence than those without a cardiac history. The effect size observed in this analysis strongly supports existing evidence linking cardiac dysfunction, arrhythmia, embolic phenomena, and systemic vascular disease to cerebrovascular risk.

The pronounced association further validates the clinical relevance of cardiovascular comorbidities as core predictive variables in stroke modeling frameworks.

8. Integrated Interpretation and Methodological Implications

The collective evidence generated across Figures 1–9 and Table 1 converged toward several clinically meaningful conclusions with direct implications for predictive modeling and translational healthcare applications.

Age emerged consistently as the dominant univariate predictor of stroke occurrence. Stroke prevalence increased substantially among older individuals, particularly beyond the sixth decade of life. This finding aligns with established biological mechanisms of vascular aging and epidemiological evidence regarding cerebrovascular disease progression.

Traditional cardiovascular risk factors, including hypertension, heart disease, and dysglycemia, demonstrated robust and clinically interpretable associations with stroke occurrence. These variables therefore warrant prioritization during feature engineering and predictive model construction.

BMI demonstrated comparatively weak predictive utility when considered independently. This observation suggests that adiposity may contribute indirectly to stroke risk through metabolic and cardiovascular pathways rather than functioning as a strong direct predictor.

The severe class imbalance present within the outcome variable necessitates careful methodological adaptation in future machine learning frameworks. Predictive modeling should incorporate imbalance-aware metrics and potentially employ techniques such as SMOTE, ADASYN, threshold optimization, or cost-sensitive learning to ensure clinically meaningful detection of stroke-positive cases.

The limited amount of missing data observed in the dataset introduces minimal analytical bias. Nonetheless, future studies may benefit from multiple imputation strategies capable of more comprehensively characterizing uncertainty associated with incomplete observations.

Because the dataset is cross-sectional, causal inference regarding temporal relationships between predictors and stroke onset cannot be definitively established. However, the observed statistical associations provide a strong empirical foundation for hypothesis generation and predictive modeling.

Several broader analytical considerations additionally warrant acknowledgment. The dataset originated from a publicly available repository, introducing potential selection biases related to healthcare access patterns, documentation practices, and population representativeness. Consequently, external validity and generalizability may be limited across geographic regions or healthcare systems.

Moreover, the absence of detailed lifestyle and behavioral variables such as physical activity, dietary habits, alcohol consumption, and sleep quality may contribute to residual confounding. Future data collection efforts should therefore prioritize richer phenotypic characterization to improve predictive performance and clinical applicability.

Finally, although the exploratory analyses presented herein establish a rigorous analytical foundation, future predictive models will require validation using temporally and geographically distinct cohorts to assess transportability, robustness, and clinical readiness.

9. Conclusion

This comprehensive exploratory investigation provides a statistically rigorous and clinically interpretable characterization of stroke-associated risk factors within a heterogeneous patient population.

The analysis demonstrates that age, cardiovascular comorbidities, and metabolic dysregulation constitute the most influential predictors associated with stroke occurrence, while BMI exhibits comparatively weaker independent discriminative value.

The integrated descriptive, correlational, and stratified analyses collectively establish a robust empirical framework for future machine learning based stroke prediction systems. Furthermore, the detailed documentation of feature distributions, class imbalance, multivariate relationships, and methodological considerations supports transparency, reproducibility, and scientific rigor.

Overall, the findings provide essential guidance for subsequent stages of feature engineering, model development, validation, and clinical translation, ultimately contributing toward the development of clinically actionable tools for early stroke risk stratification and targeted preventive intervention.

References

[1] Boehme, A. K., Esenwa, C., Elkind, M. S. (2017). Stroke risk factors, genetics, and prevention. *Circulation Research, 120*, 472–495. https://doi.org/10.1161/CIRCRESAHA.116.308398.

[2] Mozaffarian, D., Benjamin, E. J., Go, A. S., Arnett, D. K., Blaha, M. J., Cushman, M., Das, S. R., de Ferranti, S., Després, J. P., Fullerton, H. J., Howard, V. J., Huffman, M. D., Isasi, C. R., Jiménez, M. C., Judd, S. E., Kissela, B. M., Lichtman, J. H., Lisabeth, L. D., Liu, S., ... American Heart Association Statistics Committee and Stroke Statistics Subcommittee. (2016). Heart disease and stroke statistics 2016 update: A report from the American Heart Association. *Circulation, 133*(4), e38–e360. https://doi.org/10.1161/CIR.0000000000000350.

[3] Patne, S., Chintale, K. (2016). Study of clinical profile of stroke patients in rural tertiary health care centre. *International Journal of Advances in Medicine, 3*(3), 666–670. https://doi.org/10.18203/2349-3933.ijam201-62514.

[4] Katan, M., Luft, A. (2018). Global burden of stroke. *Seminars in Neurology, 38*(2), 208–211.

[5] Mendis, S., Davis, S., Norrving, B. (2015). Organizational update: The World Health Organization global status report on noncommunicable diseases 2014; One more landmark step in the combat against stroke and vascular disease. *Stroke, 46*(5), e121–e122.

[6] Lee, D. W., Jang, T. W., Kim, H. R., Kang, M. Y. (2021). The relationship between working hours and lifestyle behaviors: Evidence from a population-based panel study in Korea. *Journal of Occupational Health, 63*(1), e12280.

[7] Kang, M. Y., Cho, S. H., Yoo, M. S., Kim, T., Hong, Y. C. (2014). Long working hours may increase risk of coronary heart disease. *American Journal of Industrial Medicine, 57*(11), 1227–1234.

[8] O'Donnell, M. J., Chin, S. L., Rangarajan, S., Xavier, D., Liu, L., Zhang, H., INTERSTROKE Investigators. (2016). Global and regional effects of potentially modifiable risk factors associated with acute stroke in 32 countries (INTERSTROKE): A case-control study. *The Lancet, 388*(10046), 761–775.

[9] Murugesan, V., Natesan, M., Sulthana, V., Donapaty, P. R., Donapaty, P. (2024). Exploring factors influencing stroke risk: Insights from a predictive analysis. *Cureus, 16*(8).

[10] Meschia, J. F., Bushnell, C., Boden-Albala, B., Braun, L. T., Bravata, D. M., Chaturvedi, S., Creager, M. A., Eckel, R. H., Elkind, M. S. V., Fornage, M., Goldstein, L. B., Greenberg, S. M., Horvath, S. E., Iadecola, C., Jauch, E. C., Moore, W. S., Wilson, J. A. (2014). Guidelines for the primary prevention of stroke: A statement for healthcare professionals from the American Heart Association/American Stroke Association. *Stroke, 45*(12), 3754–3832.

[11] Harmsen, P., Lappas, G., Rosengren, A., Wilhelmsen, L. (2006). Long-term risk factors for stroke: Twenty-eight years of follow-up of 7,457 middle-aged men in Göteborg, Sweden. *Stroke, 37*(7), 1663–1667.

[12] Nwosu, C. S., Dev, S., Bhardwaj, P., Veeravalli, B., & John, D. (2019). Predicting stroke from electronic health records. *In 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society*

(EMBC) (p. 5704–5707). *IEEE*.

[13] Pathan, M. S., Jianbiao, Z., John, D., Nag, A., Dev, S. (2020). Identifying stroke indicators using rough sets. *IEEE Access, 8*, 210318–210327.

[14] Ferrario, M. M., Roncaioli, M., Veronesi, G., Holtermann, A., Clays, E., Borchini, R., Cesana, G. (2018). Differing associations for sport versus occupational physical activity and cardiovascular risk. *Heart, 104*(14), 1165–1172.

[15] Oikarinen, A., Engblom, J., Kääriäinen, M., Kyngäs, H. (2015). Risk factor-related lifestyle habits of hospital-admitted stroke patients: An exploratory study. *Journal of Clinical Nursing, 24*(15–16), 2219–2230.

[16] Dev, S., Wang, H., Nwosu, C. S., Jain, N., Veeravalli, B., John, D. (2022). A predictive analytics approach for stroke prediction using machine learning and neural networks. *Healthcare Analytics, 2*, 100032.

[17] Kaur, R., Hambarde, K., George, R., Hussain, A., Gomkar, C., Sonawani, S. (2022). Stroke prediction using optimization and exploratory data analysis. In *2022 IEEE Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI)* (p. 1–7). *IEEE*.

[18] Jeena, R. S., Kumar, S. (2016). Stroke prediction using SVM. In *2016 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT)* (pp. 600–602). *IEEE*.

[19] Hanifa, S. M., Raja S, K. (2010). Stroke risk prediction through non-linear support vector classification models. *International Journal of Advanced Research in Computer Science, 1*(3), 47–53.

[20] Chun, M., Clarke, R., Cairns, B. J., Clifton, D., Bennett, D., Chen, Y., Holmes, M. V. (2021). Stroke risk prediction using machine learning: A prospective cohort study of 0.5 million Chinese adults. *Journal of the American Medical Informatics Association, 28*(8), 1719–1727.

[21] Fu, W. (2024). Exploratory data analysis and machine learning models for stroke prediction. In *Proceedings of the International Conference on Data Analytics, Machine Learning, and IoT* (p. 211–217).

[22] Chechlacz, M., Rotshtein, P., Demeyere, N., Bickerton, W. L., Humphreys, G. W. (2014). The frequency and severity of extinction after stroke affecting different vascular territories. *Neuropsychologia, 54*, 11–17.

[23] Vallar, G., Rusconi, M. L., Bignamini, L., Geminiani, G., Perani, D. (1994). Anatomical correlates of visual and tactile extinction in humans: A clinical CT scan study. *Journal of Neurology, Neurosurgery Psychiatry, 57*(4), 464–470.

[24] Vuilleumier, P. (2013). Mapping the functional neuroanatomy of spatial neglect and human parietal lobe functions: Progress and challenges. *Annals of the New York Academy of Sciences, 1296*(1), 50–74.

[25] Becker, E., Karnath, H.-O. (2007). Incidence of visual extinction after left versus right hemisphere stroke. *Stroke, 38*(12), 3172–3174.

[26] Umarova, R. M., Saur, D., Kaller, C. P., Vry, M. S., Glauche, V., Mader, I., Weiller, C. (2011). Acute visual neglect and extinction: Distinct functional state of the visuospatial attention system. *Brain, 134*(11), 3310–3325.

[27] Vossel, S., Eschenbeck, P., Weiss, P. H., Weidner, R., Saliger, J., Karbe, H., Fink, G. R. (2011). Visual extinction in relation to visuospatial neglect after right-hemispheric stroke: Quantitative assessment and statistical lesion-symptom mapping. *Journal of Neurology, Neurosurgery & Psychiatry, 82*(8), 862–868.

[28] Kamtchum-Tatuene, J., Allali, G., Saj, A., Bernasconi, F., Vuilleumier, P. (2017). An exploratory cohort study of sensory extinction in acute stroke: Prevalence, risk factors, and time course. *Journal of Neural Transmission, 124*(4), 483–494.