

# Machine Learning Applications Within the Earlier Medicine Framework for Stroke: A Scoping Review

Ika Agustin Atika Putri<sup>a</sup>, Annisa Ristya Rahmanti<sup>a,1</sup>, Guardian Yoki Sanjaya<sup>a</sup>, Hanifah Wulandari<sup>a</sup>, Lutfan Lazuardi<sup>a</sup>, and Huan X Nguyen<sup>b</sup>

<sup>a</sup>*Department of Health Policy and Management, Faculty of Medicine, Public Health and Nursing, Universitas Gadjah Mada, Yogyakarta, Indonesia*

<sup>b</sup>*London Digital Twin Research Center, Middlesex University, London, United Kingdom*

ORCID ID: AR Rahmanti <https://orcid.org/0000-0001-9478-6267>, HX Nguyen <https://orcid.org/0000-0002-4105-2558>

**Abstract.** This scoping review explores how machine learning (ML) has been applied to stroke research within the Earlier Medicine framework, which promotes proactive and personalized care through primary (preventive care), secondary (acute care), and tertiary (chronic care) prevention levels. Following PRISMA-ScR guidelines, studies were identified from PubMed and Scopus, yielding 105 eligible studies. Most studies focused on tertiary prevention, such as predicting in-hospital mortality, complications, and functional outcomes, while fewer addressed early risk prediction or acute-phase detection. Common predictors included age, NIHSS score, glucose, stroke volume, BMI, hypertension, and diabetes, reflecting reliance on routine clinical data. Traditional ML models (e.g., logistic regression, random forest, SVM) remain dominant, although Ensemble/Hybrid and Deep Learning models have increased steadily since 2021. ML demonstrated promising predictive accuracy and clinical potential, though challenges remain in data heterogeneity, model transparency, and external validation.

**Keywords.** Stroke, Machine Learning, Earlier Medicine

## 1. Introduction

Recent advances in artificial intelligence (AI) and machine learning (ML) have enabled earlier, data-driven, and more personalized medical interventions. Stroke remains a leading causes of death and long-term disability worldwide, accounting for a substantial healthcare and economic burden [1]. Despite advances in acute treatments largely underscoring the need of predictive approaches to support prevention and early intervention.

The concept of Earlier Medicine, represents a paradigm shift from reactive to *predictive and preemptive* care [2]. It emphasizes the temporally proactive use of AI and big data to anticipate disease trajectories and enable Actionable, Accurate, Timely, and Individualized interventions across the classical continuum of prevention. In this framework, Earlier Prediction (primary prevention) target healthy individuals for risk

---

<sup>1</sup> Corresponding author : Annisa Ristya Rahmanti. Email: [annisaristya@ugm.ac.id](mailto:annisaristya@ugm.ac.id).

reduction before stroke onset or pre-disease detection through risk modeling and multimodal data integration; Earlier Detection (secondary prevention) supports timely and precise diagnosis and intervention for patient at risk or suffering from disease to optimize treatment decisions, prevent disease progression and reduce recurrence; and Earlier Prognosis (tertiary prevention) focuses on anticipating complications and disability to guide rehabilitation and long-term care strategies for quality of life improvement [2].

While numerous studies have explored AI and ML applications in stroke prediction, detection, and outcome modeling, existing literature often treats these dimensions separately, lacking integration under the unified lens of Earlier Medicine. A comprehensive synthesis is therefore required to summarize how ML technologies contribute across the three levels of prevention, identify prevailing data modalities and algorithms, and uncover translational gaps between computational models and clinical implementation. This scoping review examines the ML applications in stroke research through the perspective of Earlier Medicine. By categorizing studies into primary, secondary, and tertiary prevention, it aims to illuminate current research trends, methodological diversity, and opportunities for advancing proactive and personalized stroke care.

## 2. Methods

This scoping review followed the PRISMA-ScR guidelines [3] and was guided by the Earlier Medicine framework, which classifies studies into three domains: primary (AI detects early signs), secondary (AI predict disease progression), or tertiary (AI predicts dissabilities). Within each domain, studies were further categorized according into stroke outcome prediction, stroke risk prediction, stroke diagnosis, etiology prediction, complication or comorbidity prediction, and others [4].

A comprehensive search was conducted in PubMed and Scopus using a combination of vocabulary and free-text terms related to *stroke*, *machine learning*, *artificial intelligence*, *prediction*, *detection*, and *prognosis*. All retrieved records were imported into EndNote X9 for initial screening and deduplication. Three independent reviewers (IAAP, HW, ARR) screened titles and abstracts, followed by full-text screening to confirm eligibility. Any disagreements were resolved through discussion.

Data extraction, including Publication Year, Country, Database, Earlier Medicine Domain, Outcome Category, Stroke Type, Time to Onset, Data Type, ML Model, Key Predictors, and Performance Metrics, was conducted using a Notebook-LLM (Google NotebookLM, version 2, 2024) assisted sheet, with classifications reviewed and finalized by all authors. Data processing and visualization were performed using Python in Google Colaboratory (Google Colab, 2024 release), employing pandas (v2.2.2), NumPy (v2.0.2), Matplotlib (v3.10.0), and Seaborn (v0.13.2) libraries.

## 3. Result and Discussion

### 3.1. Study Selection

A total of 2,510 studies were retrieved from PubMed and Scopus. After removing 347 duplicates, 2,163 unique studies were screened. Following title and abstract review, 2,037 articles were excluded. The remaining 126 full-text articles were assessed and 21

articles were excluded for ineligibility. Finally, 105 studies were included in the qualitative synthesis (Figure 1).

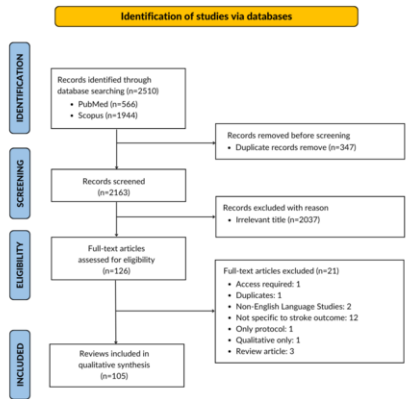


Figure 1. PRISMA Flowchart summarizing study selection

3.2. Study Characteristics and Trends in Machine Learning Applications

Across the 105 included studies, publication volume showed a steady increase over time, with a marked surge after 2021 (Figure 2a). Traditional ML methods dominated, particularly XGBoost (n = 27), random forest (n = 18), logistic regression (n = 8), gradient boosting (n = 7), and support vector machines (n = 6). Ensemble and hybrid approaches began to rise after 2023, reflecting integration of multimodal data. Deep learning was used less frequently, likely due to data size, labeling, and generalizability constraints. Reported model performance varied widely with the highest AUC observed in random forest (up to 0.9952), support vector machines (up to 0.996), and gradient boosting models (up to 0.98). However, these result were largely based on internal validation, limiting cross study comparability.

Most studies originated from China, the USA, and Europe. Although numerically balanced by income classification (46 studies from high income countries and 59 studies from LMICs), externally validated and clinically implemented models were predominantly from high-income settings, highlighting disparities in model translation.

The majority used clinical data, followed by imaging and registry data, while multimodal inputs combining clinical and imaging variables were less common. In terms of stroke classification, most studies focused on ischemic stroke (n = 64; 61%), with fewer addressing mixed populations (n = 18; 17%), unspecified types (n = 20; 19%), and only a small number on hemorrhagic stroke (n = 3; 3%). This distribution underscores the predominant reliance on ischemic stroke cohorts, likely due to their higher incidence and greater data availability in clinical repositories. Regarding time to onset, the majority of studies analyzed data from the acute phase, emphasizing early management, diagnosis, and short-term outcome prediction. A smaller proportion examined the subacute or chronic phases, typically focusing on rehabilitation and functional recovery, while pre-stroke (primary prevention) applications remained limited. This pattern indicates a continued focus on tertiary and secondary prevention domains, with underrepresentation of upstream predictive modeling for stroke prevention.

Key predictors analysis (Figure 2b) highlighted consistent emphasis on age, NIHSS score, blood glucose, lesion/volume, history of hypertension and diabetes, BMI,



while stroke risk prediction was exclusive to the primary prevention domain, focusing on pre-onset risk estimation [9]. Overall, most studies focused on tertiary prevention level, whereas primary and secondary prevention, which encompass upstream risk modeling and early detection, remain comparatively underexplored.

#### 4. Conclusion

This scoping review provides a comprehensive overview of how machine learning (ML) has been applied within the Earlier Medicine framework for stroke prevention and management. The majority of studies focused on tertiary prevention, emphasizing prediction of in-hospital outcomes, complications, and functional recovery, while fewer addressed primary and secondary prevention domains such as pre-stroke risk prediction or early detection. Across studies, traditional ML models notably logistic regression, random forest, and support vector machines, remained the most frequently applied methods, although ensemble and deep learning approaches have shown increasing adoption in recent years. Future research should prioritize upstream applications of ML for early risk stratification and pre-onset prediction, integrate multimodal real-world datasets, and adopt interpretable, clinically explainable models to enhance translational utility.

#### Acknowledgement

This research was supported by the LPDP (Ministry of Finance of the Republic of Indonesia) , The British Council ISPF Research Collaborations Programme, Grant ID #1203701580, and the Academy of Medical Sciences' ISPF Networking Grant, Grant ID: NGR1/196. ARR, GYS, HW, and LL were supported by the Community Grant Postdoctoral, FK-KMK UGM. Their financial support is gratefully acknowledged.

#### References

- [1] Institute for Health Metrics and Evaluation (IHME). *Global Burden of Disease 2021: Findings from the GBD 2021 Study*. Seattle, WA, 2024.
- [2] Iqbal U, Celi LA, Li YCJ. How can artificial intelligence make medicine more preemptive? *Journal of Medical Internet Research*; 22. Epub ahead of print 1 August 2020. DOI: 10.2196/17211.
- [3] Tricco AC, Lillie E, Zarin W, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. *Ann Intern Med* 2018; 169: 467–473.
- [4] Heo J. Application of Artificial Intelligence in Acute Ischemic Stroke: A Scoping Review. *Neurointervention* 2025; 20: 4–14.
- [5] Zuo W, Yang X. Network-based predictive models for artificial intelligence: an interpretable application of machine learning techniques in the assessment of depression in stroke patients. *BMC Geriatr* 2025; 25: 193.
- [6] Yu Y, Chen Z, Yang Y, et al. Development and validation of an interpretable machine learning model for predicting post-stroke epilepsy. *Epilepsy Res* 2024; 205: 107397.
- [7] Cao Y, Deng H, Liu S, et al. Development and validation of a machine learning-based risk prediction model for stroke-associated pneumonia in older adult hemorrhagic stroke. *Front Neurol*; 16. Epub ahead of print 18 June 2025. DOI: 10.3389/fneur.2025.1591570.
- [8] Zheng Y, Guo Z, Zhang Y, et al. Rapid triage for ischemic stroke: a machine learning-driven approach in the context of predictive, preventive and personalised medicine. *EPMA Journal* 2022; 13: 285–298.
- [9] Hassan A, Gulzar Ahmad S, Ullah Munir E, et al. Predictive modelling and identification of key risk factors for stroke using machine learning. *Sci Rep* 2024; 14: 11498.