Journal of Applied Sciences

ISSN 1812-5654

science alert

ANSInet
an open access publisher
http://ansinet.com

OPEN ACCESS Journal of Applied Sciences

ISSN 1812-5654
DOI: 10.3923/jas.2024.1.15

Review Article
Stroke Risk Factor Prediction Using Machine Learning Techniques: A Systematic Review

[1,2]Olusola Olabanjo, [3]Ashiribo Wusu, [4]Oseni Afisi and [5]Boluwaji Akinnuwesi

[1]Department of Mathematics, Morgan State University, Maryland 21251, United States of America
[2]Department of Computer Science, Lagos State University, Lagos 102101, Nigeria
[3]Department of Mathematics, Lagos State University, Lagos 102101, Nigeria
[4]Department of Philosophy, Lagos State University, Lagos 102101, Nigeria
[5]Department of Computer Science, Faculty of Science and Engineering, University of Eswatini, Kwaluseni, Eswatini

Abstract

This review addresses the global challenge of stroke, a leading cause of disability and mortality. The unpredictability and severe impact of stroke necessitate advanced prediction methods. In this work, the machine learning (ML) and deep learning (DL) techniques in stroke risk prediction were evaluated, assessing their effectiveness and application in diverse contexts. A systematic analysis of existing studies and datasets was conducted using Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA), focusing on various ML and DL algorithms used in stroke risk prediction. The 31 papers met the final inclusion criteria. The review highlights significant advancements in stroke prediction using ML and DL models, noting their ability to manage complex datasets and provide accurate predictions. However, challenges such as the need for external validation, model explainability and model transparency persist. Feature importance is further recommended to offer context-specific recommendations as stroke risk factors vary in different countries. This study also spotlights Random Forest as the outperforming model in predicting stroke risks, secondary data as the prominent dataset and China, India and Bangladesh as the country with the most stroke risk studies. The ML and DL offer promising tools for stroke risk prediction, enhancing personalized healthcare strategies. Addressing existing challenges will be crucial for their effective integration into clinical practice.

Key words: Stroke risk prediction, machine learning, deep learning, PRISMA review, predictive models, personalized medicine

Citation: Olabanjo, O., A. Wusu, O. Afisi and B. Akinnuwesi, 2024. Stroke risk factor prediction using machine learning techniques: A systematic review. J. Appl. Sci., 24: 1-15.

Corresponding Author: Olusola Olabanjo, Department of Mathematics, Morgan State University, Maryland 21251, United States of America

Competing Interest: The authors have declared that no competing interest exists.

Data Availability: All relevant data are within the paper and its supporting information files.

INTRODUCTION

Stroke, a leading cause of disability and mortality globally, is a medical condition characterized by a sudden disruption of blood supply to the brain which can have severe and often lasting effects on various functions controlled by the affected part of the brain, such as movement, speech, memory and other cognitive functions[1,2]. Stroke risk is the likelihood or probability that an individual will experience a stroke in their lifetime[3]. Stroke remains a leading cause of morbidity and mortality worldwide, emphasizing the critical need for effective risk prediction and preventive strategies. The American Heart Association's 2023 update shows that Cardiovascular Disease (CVD) remains the leading cause of death in the U.S., accounting for 928,741 deaths in 2020 and stroke was identified as the cause of 17.3% of CVD-related deaths in 2020, highlighting its significant impact. The economic burden of CVD, including stroke, is substantial, with direct and indirect costs totaling $407.3 billion between 2018 and 2019[4,5].

On a global level, according to the World Stroke Organization's Global Stroke Fact Sheet 2022, stroke remains the second-leading cause of death and the third-leading cause of death and disability combined (as measured by disability-adjusted life-years lost) globally. The estimated global cost of stroke exceeds US$721 billion, which is about 0.66% of the global GDP. From 1990 to 2019, there was a substantial increase in the burden of stroke, with a 70% increase in incident strokes, 43% increase in deaths from stroke, 102% increase in prevalent strokes and 143% increase in disability-adjusted life-years (DALYs). Much of the global stroke burden, including 86% of deaths and 89% of DALYs, is concentrated in lower-income and lower-middle-income countries[6,7]. These statistics underscore the critical importance of continued research and public health efforts in the areas of heart disease and stroke prevention, treatment and management.

Stroke risk factor prediction is crucial in the contemporary landscape of healthcare, holding profound implications for both individual patient outcomes and broader public health strategies[8,9]. The ability to foresee and identify individuals at elevated risk of stroke enables a proactive approach to healthcare, shifting the paradigm from reactive treatment to preventive care[10]. This approach is particularly salient in the context of stroke, a condition characterized by its sudden onset and potential for severe, long-lasting consequences. Accurate prediction facilitates early intervention, allowing for timely implementation of lifestyle modifications, pharmacological treatments and other preventive measures[11]. Such interventions can significantly mitigate the risk of stroke, potentially averting the onset of this life-altering event. Personalized medicine is particularly enriched by advancements in stroke risk prediction. This hinges on the understanding that the etiology and risk factors of stroke are multifaceted and individual-specific, encompassing genetic predispositions, lifestyle factors and various comorbidities. The integration of this diverse data into predictive models means healthcare can be tailored to the unique risk profile of each individual. Also, in the age of big data and artificial intelligence[12,13], the capacity to analyze vast and complex datasets offers unprecedented advancements in risk assessment, boosting a new era of predictive accuracy and healthcare customization. Additionally, insights gleaned from stroke risk prediction can guide research efforts, focusing on the most impactful areas and potentially leading to novel therapeutic and preventive strategies.

Over the past decade, the integration of machine learning (ML) and deep learning (DL) techniques into healthcare has demonstrated promising capabilities in predicting and preventing various medical conditions[14]. This systematic review aims to comprehensively evaluate the current landscape of stroke risk prediction methodologies, focusing specifically on the application of ML and DL algorithms. It shall examine the various machine learning and deep learning algorithms, assessing their effectiveness and accuracy in stroke prediction. It will also involve a thorough analysis of existing studies and datasets, comparing different techniques and models. The review of relevant literature aims to identify the most promising approaches and highlight areas for future research in stroke risk assessment using artificial intelligence techniques.

Risk factor prediction of stroke using machine learning and deep learning models: Stroke, a leading cause of disability and death globally, is influenced by a variety of risk factors, which are crucial to identify for its prevention and management. These factors are broadly divided into non-modifiable and modifiable categories[15,16]. Non-modifiable risk factors include age (with risk increasing as one ages), gender (men are more likely to have strokes, but women are more likely to die from them), ethnicity (certain ethnicities like African Americans have higher risks) and family history or genetic predispositions. These are factors that a human has no control over. On the other hand, modifiable risk factors present opportunities for intervention and prevention. These include hypertension, the most significant risk factor, heart diseases like atrial fibrillation, diabetes, high cholesterol levels and lifestyle factors such as smoking, alcohol and drug use, physical inactivity, obesity, poor diet and sleep disorders[17,18]. Psychological factors like stress and depression

also contribute indirectly, often due to associated unhealthy behaviors. Managing these risk factors involves lifestyle changes such as regular exercise, a healthy diet, avoiding tobacco and excessive alcohol and maintaining a healthy weight. For high-risk individuals, medical interventions may include medications for blood pressure, cholesterol and clot prevention.

The utilization of machine learning and deep learning models for the prediction of stroke risk factors is a significant advancement in medical science, particularly in preventive medicine and neurology. The ever-advancing field of machine learning and deep learning has ushered in a transformative era[19,20] in the prediction of stroke risk factors, offering a more scientific approach to understanding and mitigating the risk of the crippling condition. These advanced computational models and algorithms excel in their ability to sift through vast and complex datasets, including electronic health records, genetic information and lifestyle data, to unearth subtle patterns and correlations that might elude traditional statistical methods. Machine learning algorithms, with their capacity to learn from and make predictions based on data[21,22], are particularly clever at identifying individuals at high risk for stroke, enabling early and targeted interventions. Deep learning, a subset of machine learning characterized by neural networks with multiple layers, further enhances this predictive power, allowing for the analysis of incredibly intricate and layered data structures[23]. In stroke risk prediction, where the interplay of numerous risk factors demands a sophisticated analysis approach, machine learning and deep learning have proven highly effective.

The general architecture of machine learning application to risk factor analysis of stroke were shown in Fig. 1. The data collection phase is the stage where the kind of dataset to be used is determined. These datasets often include patient demographics, medical histories, lifestyle factors, genetic information and clinical parameters like blood pressure and cholesterol levels. The data then undergoes preprocessing, a critical step involving cleaning, normalizing and transforming the data into a format suitable for analysis. This often includes handling missing values, encoding categorical data, or scaling numerical values. The core of the architecture is the model development phase, where various machine learning algorithms (like decision trees, support vector machines, or logistic regression) or deep learning networks (like convolutional neural networks or recurrent neural networks) are trained on a portion of the dataset. These models learn to identify patterns and relationships within the data that are indicative of stroke risk. The trained model is then validated and tested on a separate set of data to evaluate its performance, typically using metrics such as area under curve, accuracy, sensitivity and specificity. Another important aspect of this modeling is feature selection, which involves identifying the most relevant features for the predictive model. This is essential because irrelevant or redundant features can decrease model performance. Techniques like correlation analysis, principal component analysis, or wrapper methods can be used to identify the most informative features[28].

Limitations of traditional methods for stroke risk factor prediction: The prediction of stroke risk factors has

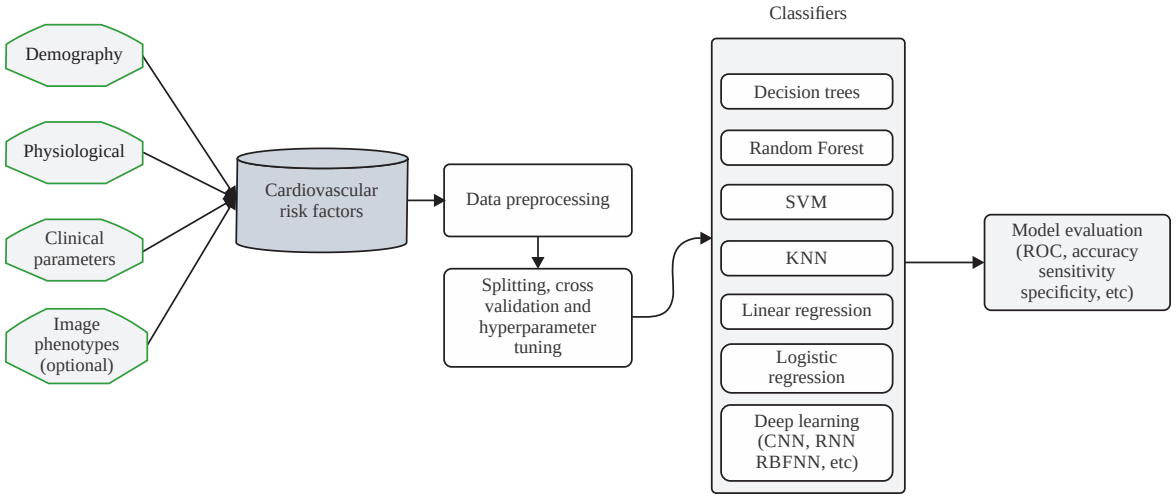

Fig. 1: Classical machine learning architecture for prediction of stroke risk factors[24-27]

traditionally been rooted in clinical assessments and standardized risk scoring systems, which form the cornerstone of preventive strategies in healthcare[29]. Among the most prominent of these is the Framingham Stroke Risk Profile, a tool developed from the Framingham Heart Study, a large, long-term, ongoing cardiovascular cohort study initiated in 1948[30]. This scoring system evaluates several key risk factors including age, blood pressure, the use of antihypertensive therapy, diabetes mellitus, cigarette smoking, prior cardiovascular disease and the presence of atrial fibrillation to estimate an individual's 10-year stroke risk. Another popular tool in medicine is the CHADS2 score, later refined into the CHA2DS2-VASc score, specifically designed to assess stroke risk in patients with atrial fibrillation. This score incorporates clinical factors such as congestive heart failure, hypertension, age, diabetes, previous stroke or transient ischemic attack, vascular disease and gender[31,32].

Despite their widespread adoption and utility in clinical settings, these traditional methods exhibit certain limitations. Primarily, they tend to focus on a restricted set of risk factors and may not fully capture the complex, multifactorial nature of stroke risk in diverse populations. The generalized nature of these models, often based on specific population cohorts, can also limit their applicability across varied ethnic and genetic backgrounds. Additionally, these conventional risk assessment tools typically employ linear statistical methods, which might not adequately represent the nonlinear interactions and relationships among multiple risk factors. This limitation is particularly pertinent in the context of emerging evidence suggesting that the interactions between lifestyle, genetic predispositions and environmental factors play a crucial role in stroke risk[33]. The advent of machine learning and deep learning offers a promising alternative, with these advanced computational techniques capable of handling high-dimensional data and discerning intricate patterns beyond the scope of traditional models. As the field of stroke research continues to evolve, there is a growing emphasis on harnessing these innovative technologies to enhance the accuracy and predictive power of stroke risk assessments, paving the way for more personalized and effective preventive healthcare strategies.

Review of related works: Many review and survey papers have investigated the application of machine learning and deep learning models to analyzing and predicting the risk factors of stroke. These papers have explored the prospects and challenges of several deep learning models on various datasets and image modalities and presented the findings of the authors in the papers. Table 1 summarizes some review works on the same subject matter being focused on in this study.

Review articles have made significant contributions to exploring the role of machine learning (ML) and deep learning (DL) models in predicting and analyzing stroke risk factors. However, this body of work has certain limitations. Firstly, the number of review articles that meet the inclusion criteria set by most authors is relatively small, especially when considering the vast volume of literature published daily. Secondly, a majority of these studies are based on single-center data, which limits the generalizability of the models. In addition, clinical validation is often not included in these studies. Moreover, many studies overlook crucial aspects such as dataset selection, model choice in ML/DL and strategies for hyperparameter tuning and optimization. Our research aims to address these gaps.

Stroke risk dataset: Stroke risk datasets play a pivotal role in machine learning (ML) for predicting the likelihood of a stroke. These datasets typically include demographic information, medical histories, lifestyle factors and biomarker data from individuals, allowing ML algorithms to uncover complex patterns and interactions among risk factors. The richness and diversity of the data are crucial for developing accurate and generalizable prediction models. Well-curated stroke risk datasets enable researchers to train, test and validate ML models, which can lead to early intervention and personalized healthcare strategies, ultimately aiming to reduce the incidence and impact of stroke. Table 2 presents some widely used publicly available datasets for prediction and analysis of stroke risk in patients.

Scope of review: This study aims to address the following research questions in the context of prediction and assessment of risk factors of stroke with ML and DL techniques. This can be utilized by researchers and medics to obtain a comprehensive view of the evolution of these techniques, datasets, modalities and the effectiveness of these techniques in the effective analysis of various types of stroke disease. The following research questions (RQs) are considered in this study:

RQ1: What are the trends and evolutions of this study?
RQ2: What ML and DL models are used for this study?
RQ3: What datasets are publicly available?
RQ4: What are the necessary considerations for application of these Artificial Intelligence (AI) techniques in stroke risk factor prediction and analysis?
RQ5: What are the limitations so far identified by authors?
RQ6: What are the future directions for this research?

Table 1: Summary of some related review works for stroke risk factor prediction using machine learning methods

Year	Articles	Summary	Reference
2023	28	This work focused on stroke mortality prediction using machine learning. It analyzed mostly retrospective studies. Authors reported that machine learning models showed a wide range of predictive accuracy (AUC 0.67-0.98) for short-term post-stroke mortality. The number of features used varied from 5 to 200, with age, BMI and NIHSS score being key predictors	Schwartz *et al.*[34]
2020	47	This work provided an extensive analysis of the application of machine learning (ML) techniques in the context of brain stroke. It highlighted the predominant use of Support Vector Machine (SVM) and Random Forest algorithms in these studies and identifies a research gap in stroke treatment	Sirsat *et al.*[35]
2023	12	This work examined the effectiveness of ML algorithms in classifying adult stroke patients. The review found no single algorithm superior for all cases due to varying input data and algorithm requirements. One noteworthy limitation is the heterogeneity (variability in study design, participant characteristics, data sources and methodologies) among included studies	Ruksakulpiwat *et al.*[36]
2020	18	This work is a systematic review evaluating the use of machine learning (ML) methods for predicting stroke outcomes using structured data. It assessed studies which focused on stroke outcomes like mortality and functional outcome, with common ML methods including random forests, support vector machines, decision trees and neural networks. The review identified limitations in the studies, such as inadequate reporting of ML methods, insufficient model descriptions for reproducibility and a lack of external validation. The need for improvements in study conduct and reporting was emphasized to enhance the application of ML in clinical practice	Wang *et al.*[37]
2023	13	This work assessed the effectiveness of machine learning in predicting stroke onset time. It includes a meta-analysis involving 55 machine learning models. Limitations identified included heterogeneity across studies, mainly single-center studies, which might affect generalizability and a small number of included studies	Feng *et al.*[38]
2023	286	Study reviewed the use of deep learning (DL) techniques for Cardiovascular Disease (CVD) and stroke risk stratification. It emphasized solo deep learning (SDL) and hybrid deep learning (HDL) architectures for risk assessment. The review discussed the role of DL in CVD/stroke risk stratification and notes the increasing adoption of ensemble-based DL techniques due to their reliability and accuracy	Bhagawati *et al.*[39]
2020	-	Work reviewed the development of predictive CVD risk models, discussing conventional models and their limitations, such as oversimplifying complex associations and limited applicability across different ethnicities. The review also emphasized integrating noninvasive imaging with AI for improved risk prediction. It highlights the potential of AI in enhancing CVD risk assessment but also notes the infancy of these systems and the need for further development and validation	Jamthikar *et al.*[40]

Table 2: An overview of some publicly available stroke risk factor dataset

Dataset	Description	Usage in literature
Data.world (https://data.world/datasets/stroke)	Robust database hosts a variety of stroke-related datasets. It contains datasets for a range of stroke-related data, including behavioral risk factors and prevention strategies	Revathi *et al.*[41], Kao *et al.*[42] and Naz and Ahuja[43]
Annotated clinical MRIs and metadata (https://www.nature.com/articles/sdata201811)	Dataset includes annotated clinical MRIs and detailed metadata of patients with acute stroke. It categorizes lesions as ischemic, hemorrhage, or not visible and includes demographic and clinical information recorded at admission and discharge	Liu *et al.*[44]
ATLAS (Anatomical tracings of lesions after stroke available at https://www.nature.com/articles/sdata 201811)	ATLAS dataset comprises 304 T1-weighted MRIs with manually segmented diverse lesions and metadata. It was collected from 11 cohorts worldwide and includes detailed descriptions of the type of stroke, primary lesion location, vascular territory and intensity of white matter disease. This dataset is particularly valuable for research in stroke rehabilitation and MRI-based lesion segmentation	Liew *et al.*[45]
Kaggle's stroke prediction dataset (https://www.kaggle.com/datasets/fedesoriano/ stroke-prediction-dataset)	Kaggle offers a stroke prediction dataset that is often used for machine learning and predictive modeling in stroke research. This dataset typically includes various clinical features that are predictive of stroke events	Liu *et al.*[46], Sailasya and Kumari[47] and Biswas *et al.*[48]
Nationwide registry-based cohort study for 30-day mortality prediction (Github)	A large dataset containing information on 488,947 patients, with a focus on predicting 30-day mortality after stroke, is available on GitHub. It includes a wide range of clinical and demographic data, such as age, prevalence of congestive heart failure, Atrial Fibrillation (AF), previous stroke/TIA and more. This dataset is particularly useful for developing and validating machine learning models for mortality risk stratification	Wang *et al.*[49] and Wagner *et al.*[50]

We also investigated the verifiability of these studies by checking whether a medic or radiologist was one of the contributors or the results of the model was stated to have been externally validated by one.

METHODS

This review article explores, evaluates and draws conclusion from various studies on predicting stroke risk factors using machine learning (ML) techniques. The aim is to offer a thorough overview of the topic, encapsulating diverse ML methodologies, datasets, models and a range of optimization strategies for training these models. The authors will engage in comparative analyses, address challenges and limitations encountered and propose potential avenues for future research and enhancement. The methodology for this review adheres to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines.

Database search and eligibility criteria: In this systematic review, we developed a search strategy to explore Google Scholar for relevant, up-to-date research publications on the use of ML models in clinically predicting stroke risk factors. Other databases like ResearchGate were used as a secondary resource for preliminary and expository discussions. The investigation timeframe spans from 2017 to 2023. These sources were chosen due to their extensive research publication indexing in this area.

Review strategy: The methodology for this review encompasses various stages, including study selection, defining the research design, formulating a search strategy, identifying information sources and outlining data collection methods. We assessed papers that met our initial inclusion and exclusion criteria. Exclusions were made for editorials, commentaries, letters, preprints, databases outside of the four primary categories and other types of manuscripts. Our search strategy involved: (a) Creating search terms by pinpointing key keywords, necessary actions and anticipated outcomes; (b) Identifying synonyms or alternative terms for these keywords, (c) Setting specific exclusion parameters for the search and (d) Using Boolean operators to structure the search query effectively.

Results of (a): Machine learning, stroke, risk factors, prediction and diagnosis

Results of (b): Prediction/diagnosis/classification, machine learning, stroke and risk factors

Results of (c): Review, systematic review, preprint, unrelated risk factors, treatment, MRI and CT scan

Results of (d): a, b, c combined using AND OR

Publications were selected from peer-reviewed literature using the generated search phrase on Google Scholar. Conference proceedings, journals, book chapters and whole books were vetted. The initial number of results returned was 1022; of those, 986 met the initial selection criterion and 31 fulfilled the final requirements. The studies were appropriately grouped. A PRISMA flowchart for study selection was utilized following Fig. 2.

Characteristics of studies: The characteristics of the 31 reviewed articles were depicted in Fig. 3 showing the year distribution of included papers. Figure 3 make it clear that only recent papers were given the most priority.

Quality assessment: Most studies failed to meet at least one of the six quality criteria. Common issues included limited sample size, inadequate scientific strategies and failure to disclose results for computational techniques, impacting study quality.

Data sources and search strategy: We searched the selected databases for studies published before December, 2023 but not earlier than 2017. Keywords were searched in subject headings, titles or abstracts using Boolean operators, with the language restricted to English. Reference lists of primary studies and review articles were also reviewed.

Inclusion and exclusion criteria: Publications which applied machine learning (ML) and deep learning (DL) to predict and analyze stroke risk factors were reviewed in this research. To meet our inclusion criteria, papers needed to detail the Artificial Intelligence (AI) methods employed and the specific aspects of stroke risk they examined. Works focusing on crucial datasets and their analysis methods were also considered. Excluded from this review were preprints, publications from non-selected databases, opinion pieces, commentaries, non-English language articles, editorials, narrative reviews, case reports, conference abstracts and duplicated studies. Additionally, articles with redundant techniques and findings were not included in the analysis.

Data extraction: Full texts of the included papers were obtained and reviewers independently gathered data from each study, resolving any discrepancies through mutual agreement. The information extracted encompassed references, year of publication, context of the study, the machine learning methods used, details of the dataset and

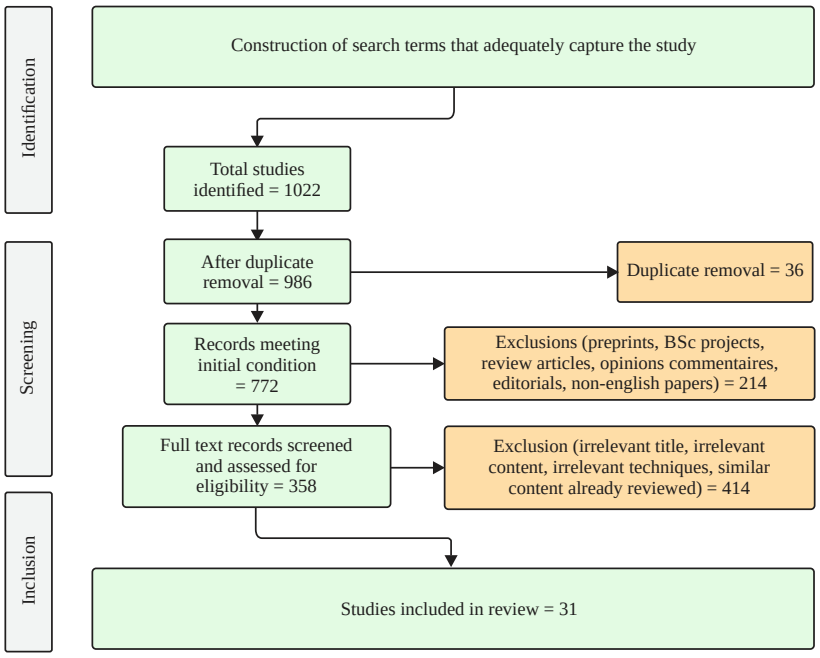

Fig. 2: PRISMA-Scr numerical flow guideline for systematic review employed in this study

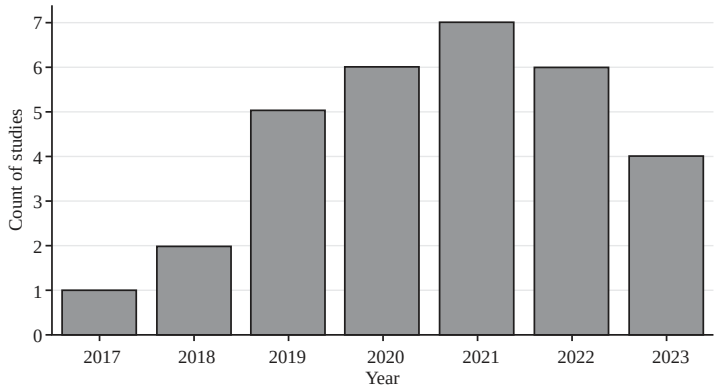

Fig. 3: Distribution of included studies by year of publication

any imaging techniques employed, performance metrics and the accuracy attained. A comparative analysis was then performed using the data compiled from these sources.

Data synthesis: Included papers were examined in terms of model types, datasets, preprocessing techniques, features extracted and reported performance metrics. Performance metrics of interest included sensitivity, specificity, accuracy and the Area under the Receiver Operating Characteristic Curve (AUC-ROC).

Risk of bias assessment: The evaluation focused on analyzing the methodological quality and identifying possible sources of bias that might affect the credibility of the results. Biases could arise from factors such as the use of datasets from a single center or imbalanced class distributions, which might impact the generalizability of the models. Furthermore, issues like insufficient documentation of data preprocessing steps, problems in model fitting and inadequate details on hyperparameter tuning could pose challenges to the reproducibility of the findings.

RESULTS

Table 3 shows the results of this review process for all included papers. It summarizes the studies by stating column-wise the objectives, year and country of study, type of dataset collected from a clinical setting or obtained from a

publicly available repository, ML models used and metrics reported, external validation (EV) reported, including the strengths and weaknesses of the study.

Table 3 presents a comprehensive review of various studies on machine learning for stroke risk factor analysis. These studies, conducted between 2017 and 2023, primarily focused on predicting stroke occurrence, its types and associated risks using a range of machine learning models like Decision Trees, Naive Bayes, Random Forest and Artificial Neural Networks. Most studies leverage large datasets from healthcare centers and employ metrics such as accuracy, precision and AUC for model evaluation. A common strength across these studies is the innovative application of machine learning in stroke prediction and risk assessment, often integrating multiple data types and machine learning techniques. However, a recurring limitation is the lack of external validation and potential biases in datasets, which could affect the generalizability and accuracy of the predictive models. The synthesis of these studies highlights the evolving role of machine learning in healthcare, specifically in enhancing stroke risk prediction and management.

It is also noteworthy as evident in Fig. 4 that random forest is the most popular with the highest frequency of occurrence and performance in the review papers. This is followed by support vector machine and decision trees. Linear regression and Naïve Bayes also showed promising potential in identifying and predicting risk factors of stroke. Figure 5 shows the country distribution of included and spotlights China, India and Bangladesh as the countries with the highest stroke studies.

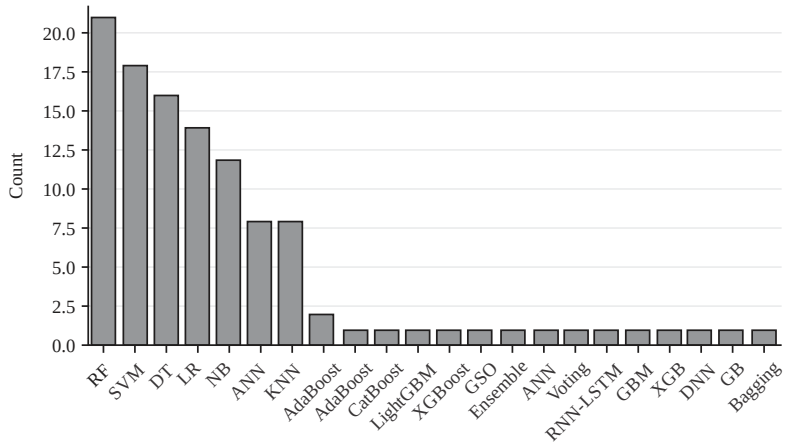

Fig. 4: Frequency of occurrence of ML models

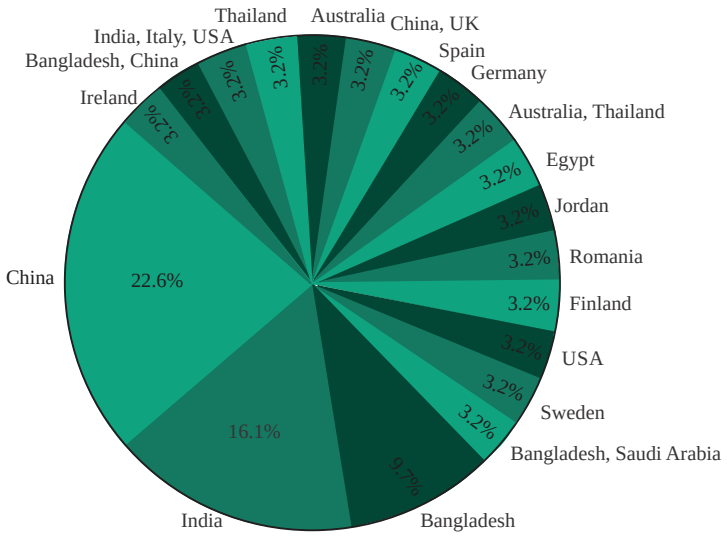

Fig. 5: Distribution of studies by country

Table 3: Summary of included studies

Study objective (s)	Year	Country	Dataset	Models used	Metrics used	EV	Strengths	Weaknesses	References
Authors highlighted the use of machine learning algorithms for this purpose and analyzes their performance, as well as identified significant features from datasets for stroke risk prediction	2020	Bangladesh	Secondary: Of 62001×12 dimension	LR, RF and DT	98% accuracy and precision, 99% recall and FI-score	No	Large dataset used which implies that the model has enough training set to generalize solutions. Feature selection was implemented before prediction	No validation reported. Also deep learning models will fit and generalize the solution better than linear models	Shafiul Azam *et al.*[51]
Authors developed a machine learning model for cardiovascular/stroke risk prediction by integrating carotid ultrasound image-based phenotypes (CUSIP) with conventional risk factors (CRF)	2020	India, Italy and United States of America	Primary: Image+other risk factors	RF	99% AUC	No	The integration of image-based phenotypes with traditional risk factors represents a novel approach in the stroke risk assessment study	Possible slight bias in the overall estimation of predicted risk due to the use surrogate image-based biomarkers. Deep learning models would also perform better due to their image analysis strength	Jamthikar *et al.*[52]
Authors developed a stroke risk prediction model using a novel Hybrid Deep Transfer Learning-based Stroke Risk Prediction (HDTL-SRP) framework by utilizing the knowledge structure from multiple correlated sources, such as external stroke data and chronic diseases data like hypertension and diabetes	2021	China	Primary: Electronic Health Record (EHR) databases of three hospitals	DNN	76% accuracy, 72% recall, 76% F1 score and 83% AUC	Yes	Multiple data sources, large size of dataset, external validation with real-world dataset	This study may be challenged in the integration and interpretation of data from varied sources	Chen *et al.*[53]
A machine learning-based model for prognosis prediction in AIS patients was developed using data collected from the Second Affiliated Hospital of Xuzhou Medical University between August, 2017 and July, 2019	2023	China	Primary: Variable including neuron specific enolase (NSE), Homocysteine (HCY), S-100β, dysphagia, C-Reactive Protein (CRP) and anticoagulation	NB, XGB, RF, DT, GBM and LR	RF outperformed others with AUC of 93%, accuracy of 79%, sensitivity of 76%, specificity of 88% and F1 score of 70%	Yes-using external cohort	Comprehensive analysis using both univariate and multivariate logistic regression to identify key prognostic factors Development of an online tool based on the predictive model to assist clinicians in optimizing patient treatment	Potential biases due to the retrospective nature of the study limited generalizability as the data were collected from a single medical center	Wang *et al.*[54]
To predict the occurrence of stroke using various machine learning algorithms, based on physiological factors	2021	India	Secondary: Kaggle (5110×12)	LR, DT, RF, KNN, SVM and NB	NB performed best with 82% accuracy Precision, recall and F1 were also reported	No	Diverse physiological factors for stroke prediction and development of a web application for easy user interaction and stroke risk assessment	Dataset used was highly imbalanced, which could affect the predictive performance of proposed model	Sailasya and Kumari[47]
To predict mortality and cerebrovascular events in mitral regurgitation patients using a gradient boosting machine (GBM) model, considering comorbidities, P-wave and echocardiographic measurements	2023	China and United Kingdom	Secondary-706 patients	LR, DT, RF, SVM and ANN	No numerical performance metric reported	No	Comprehensive analysis considering comorbidities and echocardiographic measurements is a major relative strength	The study is based on data from a single tertiary center, which may limit the generalizability of the findings to other populations	Zhou *et al.*[55]
To predict intrahospital clinical outcomes in stroke patients undergoing Transcatheter Aortic Valve Implantation (TAVI)	2021	Germany	Primary-451 patients	ANN, SVM and RF	AUC of 97%	No	The study employed neural networks, support vector machines and random forests, evaluated using five -fold nested cross-validation	Dataset used was small and not externally validated	Gomes *et al.*[56]
To predict stroke using Logistic Regression, Decision Tree, Random Forest and Voting Classifier	2021	Bangladesh, Saudi Arabia	Secondary: 5110×12 dimension	RF, DT, Voting and LR	RF did best with 96% accuracy	No	Study included comparison with previous studies; high accuracy demonstrated by models, especially Random Forest	Dataset used was small and not externally validated	Tazin *et al.*[57]
To predict brain strokes using machine learning classifiers and a stacking ensemble classifier	2022	Egypt	Secondary-706 patients	RF, KNN, LR, SVM and NB	KNN did best with 97% accuracy	No	A stacking ensemble classifier integrating various classifiers for enhanced prediction accuracy	No external validation and feature importance analysis was not performed	Mostafa *et al.*[58]

Table 3: Continued

Study objective (s)	Year	Country	Dataset	Models used	Metrics used	EV	Strengths	Weaknesses	References
To select features important to stroke prognosis using an integrated machine learning approach on the International Stroke Trial dataset	2020	China	Secondary	SVM, RF, AdaBoost and NB	F1-79%	No	The use of advanced machine learning algorithms for feature selection and prognosis prediction	No external validation or other standard metric reported because prediction was not done	Fang *et al.*[59]
To predict ischemic stroke outcomes post intra-arterial therapy using machine learning algorithms	2020	India	Secondary	ANN and SVM	98% accuracy	No	Relatively high accuracy	Requires large training datasets for improved performance; potential issues with data accuracy	Pathanjali *et al.*[60]
To predict the likelihood of a person having a stroke based on medical risk factors such as smoking status, heart disease, glucose value and hypertension	2022	India	Secondary	SVM, RF and KNN	94.6%	No	Utilization of multiple models for stroke prediction, potentially improving system performance	Details on dataset not well articulated Result not externally validated	Telu and Padimi[61]
To investigate the main risk factors for stroke in Shanxi Province, China, using machine learning models	2021	China	Primary: Of 27,583 residents from 2017 to 2020	DT, RF and LR	RF outperformed others with 80% precision, 85% recall and 85% accuracy	No	Utilization of a large dataset and multiple models for a comprehensive analysis	China-based studies means the result must be externally validated on new datasets to verify the validity and acceptability of the model	Liu *et al.*[62]
To develop a feature selection and classification model for detecting stroke risk using clinical data	2018	China	Primary	SVM and GSO	83% accuracy	No	A novel hybrid feature selection model combining Support Vector Machines (SVM) with Glow-Worm Swarm Optimization (GSO), enhancing prediction accuracy	External validation was not reported	Zhang *et al.*[63]
To identify novel, previously unidentified risk factors for stroke using a machine learning model	2022	USA	Secondary-Kaggle-5110×12 dimension	ANN	92% accuracy and 73% AUC	No	Utilization of an artificial neural network model for analysis; identification of significant risk factors like occupation	Secondary data, no external validation, deep learning can perform better in instances where ANN is preferred	Keerthy[64]
To predict stroke among the elderly in China using machine learning methods, particularly in the context of imbalanced data	2020	China	Secondary-Chinese Longitudinal Healthy Longevity study	LR, SVM and RF	AUC-78%	No	Incorporates various machine learning methods and data balancing techniques to address imbalanced datasets	Limited discussion on the potential limitations of the study, such as biases in self-reported data	Wu and Fang[65]

DT: Decision tree, NB: Naive Bayes, ANN: Artificial Neural Network, KNN: K-Nearest Neighbors, LR: Logistic regression, SVM: Support Vector Machine, RF: Random forest, GB: Gradient boosting, DNN: Deep neural network, XGB: Extreme gradient boosting, GBM: Gradient boosting machine, RNN-LSTM: Recurrent Neural Network-Long Short Term Memory, Voting: Voting classifier, AdaBoost: Adaptive boosting, GSO: Group search optimizer, XGBoost: eXtreme Gradient Boosting, LightGBM: Light Gradient Boosting Machine, CatBoost: Categorical boosting and Bagging: Bootstrap aggregating

DISCUSSION

This study presents an in-depth analysis of current research on machine learning (ML) applications in stroke risk prediction. It covers studies which focused on the potential of ML models to identify and weigh various risk factors, offering a predictive insight that could revolutionize stroke prevention strategies. Through this systematic review, the study highlights the recurring challenge of limited external validation, which is essential for verifying model performance across diverse populations. It also notes the critical role of large, varied datasets in developing models with higher generalizability. A notable finding is the methodological diversity among the analyzed studies, with many employing advanced algorithms like deep learning. However, issues such as small sample sizes and inconsistent reporting standards underscore the need for a more standardized approach in future research. This work also suggests that while ML techniques hold promise, there is a pressing need for transparent and interpretable ML models. Such clarity would allow healthcare professionals to integrate ML-assisted predictions into clinical practice effectively, enhancing patient outcomes through personalized risk assessment. This summary encapsulates the key messages of the review while adhering to the concise summary constraints.

There are gaps identified in the examined papers, however. Firstly, a recurring limitation is the lack of external validation in many of these studies. External validation involves testing the ML/DL models on a completely independent dataset, ideally from a different demographic or geographic population than the one used for training the model. This step is crucial for assessing the generalizability and applicability of the models to diverse patient populations. Without external validation, the predictive power of these models may be overestimated and their utility in real-world clinical settings remains uncertain.

Additionally, many studies suffer from small sample sizes and the use of data from single centers. This can lead to models that are overly tailored to specific patient populations, reducing their effectiveness in broader healthcare contexts. Furthermore, the studies often lack diversity in terms of ethnicity and socio-economic backgrounds of the patients, which is vital for developing models that are universally applicable.

Another significant gap is the transparency and interpretability of ML/DL models. Many studies do not provide sufficient detail on the features used, the model architecture and the decision-making process of the algorithms. This lack of transparency can hinder the acceptance and trust of these tools among healthcare practitioners, who often require a clear understanding of how decisions are made for patient care. Moreover, there is a need for more rigorous methodological approaches, including standardized reporting of model development and performance metrics. This standardization is essential for comparing results across different studies and for the advancement of the field. Finally, most studies focus primarily on predictive accuracy, often neglecting other important aspects such as model robustness, scalability and the practical feasibility of integrating these models into existing healthcare systems. Addressing these limitations is crucial for the successful translation of ML/DL research into effective, real-world clinical tools for stroke and cardiovascular risk assessment.

The availability of publicly accessible datasets for stroke risk prediction using machine learning (ML) is crucial for several reasons. First, it allows for the development and testing of predictive models across a wide range of demographic and geographic populations, ensuring the models' applicability and accuracy in diverse settings. Secondly, public datasets facilitate collaborative research, enabling scientists and healthcare professionals to share insights and improvements. This collaborative environment can lead to more rapid advancements in stroke prediction methodologies. Lastly, transparent and accessible data can enhance the reliability and trustworthiness of ML models, as external researchers can validate and scrutinize these models, ensuring their robustness and clinical relevance.

Random Forest algorithm[66] has been applied mostly in the prediction of stroke risk factors[67], understandably due to its high accuracy and ability to manage complex, large datasets. This algorithm effectively deals with both numerical and categorical data, making it suitable for diverse medical datasets often encountered in stroke research[67,68]. Its resistance to overfitting is particularly valuable, ensuring that the models developed are robust and reliable[69]. Furthermore, the capability of Random Forest to handle missing data and provide insights into variable importance makes it an indispensable tool in the nuanced field of stroke risk prediction, where understanding and weighting various risk factors accurately is crucial.

Explainability, without which there is no accountability and trustworthiness in machine learning models, is a critical aspect, focusing on making the models' decision-making processes transparent and understandable[70,71]. It is essential for clinicians and patients to trust and effectively use these predictive models. Explainable models help in identifying key risk factors contributing to stroke, enabling targeted interventions. Moreover, explainability aids in model validation and error analysis, ensuring accuracy and reliability in a typical black box model. In a domain as sensitive as

healthcare, where decisions significantly impact patient outcomes, the ability to explain and interpret model predictions is invaluable for gaining clinical acceptance and enhancing patient care[72].

This systematic review shows the potential of machine learning (ML) and deep learning (DL) techniques in enhancing stroke risk prediction and management. While these innovative models offer significant advancements in identifying and analyzing complex patterns in stroke-related data, several challenges remain. Key issues include the need for external validation to ensure the models' generalizability across diverse populations, addressing small sample sizes and improving the transparency and interpretability of ML/DL models. Addressing these challenges is crucial for the effective integration of these models into clinical practice, which could lead to more personalized and impactful healthcare strategies for stroke prevention and management.

CONCLUSION

This review endeavor comprehensively examined the utilization of machine learning (ML) and deep learning (DL) in predicting stroke risk factors, revealing significant advancements in the field. Our analysis shows the effectiveness of these models in handling complex datasets and improving prediction accuracy, essentially for personalized healthcare. However, challenges such as the need for external validation and model transparency persist. Future directions should focus on enhancing model explainability and expanding research to include diverse datasets, ensuring broader applicability and integration into clinical practice.

SIGNIFICANCE STATEMENT

This work considers the pressing global issue of stroke, a major cause of disability and mortality. Authors systematically evaluated the application of machine learning (ML) and deep learning (DL) techniques in predicting stroke risk, a critical step towards proactive healthcare using the standard PRISMA strategy. Key findings demonstrated that ML and DL models, especially Random Forest, effectively manage complex datasets and enhance predictive accuracy, which is vital for personalized healthcare strategies. The research suggests the need for more external validation, model explainability and transparency to improve clinical integration. Future research could focus on feature importance to provide tailored stroke risk assessments in diverse populations.

REFERENCES

1. Murphy, S.J.X. and D.J. Werring, 2020. Stroke: Causes and clinical features. Medicine, 48: 561-566.
2. Maida, C.D., R.L. Norrito, M. Daidone, A. Tuttolomondo and A. Pinto, 2020. Neuroinflammatory mechanisms in ischemic stroke: Focus on cardioembolic stroke, background, and therapeutic approaches. Int. J. Mol. Sci., Vol. 21. 10.3390/ijms21186454.
3. South, K., L. McCulloch, B.W. McColl, M.S.V. Elkind, S.M. Allan and C.J. Smith, 2020. Preceding infection and risk of stroke: An old concept revived by the COVID-19 pandemic. Int. J. Stroke, 15: 722-732.
4. Thayabaranathan, T., J. Kim, D.A. Cadilhac, A.G. Thrift and G.A. Donnan *et al*., 2022. Global stroke statistics 2022. Int. J. Stroke, 17: 946-956.
5. Tsao, C.W., A.W. Aday, Z.I. Almarzooq, C.A.M. Anderson and P. Arora *et al*., 2023. Heart disease and stroke statistics-2023 update: A report from the American heart association. Circulation, 147: e93-e621.
6. Feigin, V.L., M. Brainin, B. Norrving, S. Martins and R.L. Sacco *et al*., 2022. World Stroke Organization (WSO): Global stroke fact sheet 2022. Int. J. Stroke, 17: 18-29.
7. Tsao, C.W., A.W. Aday, Z.I. Almarzooq, A. Alonso and A.Z. Beaton *et al*., 2022. Heart disease and stroke statistics-2022 update: A report from the American Heart Association. Circulation, 145: e153-e639.
8. Ovbiagele, B., L.B. Goldstein, R.T. Higashida, V.J. Howard and S.C. Johnston *et al*., 2013. Forecasting the future of stroke in the United States: A policy statement from the American Heart Association and American Stroke Association. Stroke, 44: 2361-2375.
9. Saceleanu, V.M., C. Toader, H. Ples, R.A. Covache-Busuioc and H.P. C ostin *et al*., 2023. Integrative approaches in acute ischemic stroke: From symptom recognition to future innovations. Biomedicines, Vol. 11. 10.3390/biomedicines11102617.
10. Rasool, S., A. Husnain, A. Saeed, A.Y. Gill and H.K. Hussain, 2023. Harnessing predictive power: Exploring the crucial role of machine learning in early disease detection. JURIHUM: J. Inovasi Humaniora, 1: 302-315.
11. Arafa, A., Y. Kokubo, H.A. Sheerah, Y. Sakai and E. Watanabe *et al*., 2022. Developing a stroke risk prediction model using cardiovascular risk factors: The suita study. Cerebrovasc. Dis., 51: 323-330.
12. Orfanoudaki, A., E. Chesley, C. Cadisch, B. Stein, A. Nouh, M.J. Alberts and D. Bertsimas, 2020. Machine learning provides evidence that stroke risk is not linear: The non-linear Framingham stroke risk score. PLoS ONE, Vol. 15. 10.1371/journal.pone.0232414.

13. Ibrahim, M.S. and S. Saber, 2023. Machine learning and predictive analytics: Advancing disease prevention in healthcare. J. Contemp. Healthcare Anal., 7: 53-71.
14. Deo, R.C., 2015. Machine learning in medicine. Circulation, 132: 1920-1930.
15. Saini, N.R. and A.L. Gurvendra, 2022. Stroke-related risk factors: A review. Asian Pac. J. Health Sci., 9: 102-107.
16. Sacco, R.L., 1997. Risk factors, outcomes, and stroke subtypes for ischemic stroke. Neurology, 49: S39-S44.
17. Konduru, S.S.T., A. Ranjan, A. Bollisetty and V. Yadla, 2018. Assessment of risk factors I nfluencing functional outcomes in cerebral stroke patients using modified Rankin scale. World J. Pharm. Pharm. Sci., 7: 755-769.
18. Nakibuuka, J., M. Sajatovic, J. Nankabirwa, A.J. Furlan and J. Kayima *et al*., 2015. Stroke-risk factors differ between rural and urban communities: Population survey in Central Uganda. Neuroepidemiology, 44: 156-165.
19. Fashoto, S.G., B. Akinnuwesi, O. Owolabi and D. Adelekan, 2016. Decision support model for supplier selection in healthcare service delivery using analytical hierarchy process and artificial neural network. Afr. J. Bus. Manage., 10: 209-232.
20. Olabanjo, O., A. Wusu, M. Asokere, O. Afisi and B. Okugbesan *et al*., 2023. Application of machine learning and deep learning models in prostate cancer diagnosis using medical images: A systematic review. Analytics, 2: 708-744.
21. Olabanjo, O.A., A.S. Wusu and M. Manuel, 2022. A machine learning prediction of academic performance of secondary school students using radial basis function neural network. Trends Neurosci. Educ., Vol. 29. 10.1016/j.tine.2022.100190.
22. Akinnuwesi, B.A., B.O. Macaulay and B.S. Aribisala, 2020. Breast cancer risk assessment and early diagnosis using Principal Component Analysis and support vector machine techniques. Inf. Med. Unlocked, Vol. 21. 10.1016/j.imu.2020.100459.
23. Aribisala, B., O. Odusanya, O. Olabanjo, E. Wahab, O. Atilola and A. Saheed, 2022. Development of an artificial neural network model for detection of COVID-19. Int. J. Sci. Adv., 3: 377-385.
24. Dritsas, E. and M. Trigka, 2022. Stroke risk prediction with machine learning techniques. Sensors, Vol. 22. 10.3390/s22134670.
25. Chahine, Y., M.J. Magoon, B. Maidu, J.C. del Álamo, P.M. Boyle and N. Akoum, 2023. Machine learning and the conundrum of stroke risk prediction. Arrhythmia Electrophysiol. Rev., Vol. 12. 10.15420/aer.2022.34.
26. Abedi, V., V. Avula, D. Chaudhary, S. Shahjouei and A. Khan *et al*., 2021. Prediction of long-term stroke recurrence using machine learning models. J. Clin. Med., Vol. 10. 10.3390/jcm10061286.
27. Dimopoulos, A.C., M. Nikolaidou, F.F. Caballero, W. Engchuan and A. Sanchez-Niubo *et al*., 2018. Machine learning methodologies versus cardiovascular risk scores, in predicting disease risk. BMC Med. Res. Methodol., Vol. 18. 10.1186/s12874-018-0644-1.
28. Zhang, Y., Y. Zhou, D. Zhang and W. Song, 2019. A stroke risk detection: Improving hybrid feature selection method. J. Med. Internet Res., Vol. 21. 10.2196/12437.
29. Chao, T.F. and S.A. Chen, 2014. Stroke risk predictor scoring systems in atrial fibrillation. J. Atrial Fibrillation, 6: 59-63.
30. D'Agostino, R.B., P.A. Wolf, A.J. Belanger and W.B. Kannel, 1994. Stroke risk profile: Adjustment for antihypertensive medication. The framingham study. Stroke, 25: 40-43.
31. Gage, B.F., A.D. Waterman, W. Shannon, M. Boechler, M.W. Rich and M.J. Radford, 2001. Validation of clinical classification schemes for predicting stroke: Results from the National Registry of Atrial Fibrillation. JAMA, 285: 2864-2870.
32. Chen, J.Y., A.D. Zhang, H.Y. Lu, J. Guo, F.F. Wang and Z.C. Li, 2013. CHADS2 versus CHA2DS2-VASc score in assessing the stroke and throm-boembolism risk stratification in patients with atrial fibrillation: A systematic review and meta-analysis. J. Geriatric Cardiol., 10: 258-266.
33. Meschia, J.F., C. Bushnell, B. Boden-Albala, L.T. Braun and D.M. Bravata *et al*., 2014. Guidelines for the primary prevention of stroke: A statement for healthcare professionals from the American Heart Association/American Stroke Association. Stroke, 45: 3754-3832.
34. Schwartz, L., R. Anteby, E. Klang and S. Soffer, 2023. Stroke mortality prediction using machine learning: Systematic review. J. Neurol. Sci., Vol. 444. 10.1016/j.jns.2022.120529.
35. Sirsat, M.S., E. Fermé and J. Câmara, 2020. Machine learning for brain stroke: A review. J. Stroke Cerebrovascular Dis., Vol. 29. 10.1016/j.jstrokecerebrovasdis.2020.105162.
36. Ruksakulpiwat, S., W. Thongking, W. Zhou, C. Benjasirisan, L. Phianhasin, N.K. Schiltz and S. Brahmbhatt, 2023. Machine learning-based patient classification system for adults with stroke: A systematic review. Chron. Illness, 19: 26-39.
37. Wang, W., M. Kiik, N. Peek, V. Curcin and I.J. Marshall *et al*., 2020. A systematic review of machine learning models for predicting outcomes of stroke with structured data. PLoS ONE, Vol. 15. 10.1371/journal.pone.0234722.
38. Feng, J., Q. Zhang, F. Wu, J. Peng, Z. Li and Z. Chen, 2023. The value of applying machine learning in predicting the time of symptom onset in stroke patients: Systematic review and meta-analysis. J. Med. Internet Res., Vol. 25. 10.2196/44895.
39. Bhagawati, M., S. Paul, S. Agarwal, A. Protogeron and P.P. Sfikakis *et al*., 2023. Cardiovascular disease/stroke risk stratification in deep learning framework: A review. Cardiovasc. Diagn. Ther., 13: 557-598.
40. Jamthikar, A.D., D. Gupta, L. Saba, N.N. Khanna and K. Viskovic *et al*., 2020. Artificial intelligence framework for predictive cardiovascular and stroke risk assessment models: A narrative review of integrated approaches using carotid ultrasound. Comput. Biol. Med., Vol. 126. 10.1016/j.compbiomed.2020.104043.

41. Revathi, A., R. Kaladevi, K. Ramana, R.H. Jhaveri, M.R. Kumar and M.S.P. Kumar, 2022. Early detection of cognitive decline using machine learning algorithm and cognitive ability test. Secur. Commun. Networks, Vol. 2022. 10.1155/2022/4190023.
42. Kao, Y.T., C.Y. Huang, Y.A. Fang, J.C. Liu and T.H. Chang, 2023. Machine learning-based prediction of atrial fibrillation risk using electronic medical records in older aged patients. Am. J. Cardiol., 198: 56-63.
43. Naz, H. and S. Ahuja, 2020. Deep learning approach for diabetes prediction using PIMA Indian dataset. J. Diabetes Metab. Disord., 19: 391-403.
44. Liu, C.F., R. Leigh, B. Johnson, V. Urrutia and J. Hsu *et al*., 2023. A large public dataset of annotated clinical MRIs and metadata of patients with acute stroke. Sci. Data, Vol. 10. 10.1038/s41597-023-02457-9.
45. Liew, S.L., J.M. Anglin, N.W. Banks, M. Sondag and K.L. Ito *et al*., 2018. A large, open source dataset of stroke anatomical brain images and manual lesion segmentations. Sci. Data, Vol. 5. 10.1038/sdata.2018.11.
46. Liu, T., W. Fan and C. Wu, 2019. A hybrid machine learning approach to cerebral stroke prediction based on imbalanced medical dataset. Artif. Intell. Med., Vol. 101. 10.1016/j.artmed.2019.101723.
47. Sailasya, G. and G.L.A. Kumari, 2021. Analyzing the performance of stroke prediction using ML classification algorithms. Int. J. Adv. Comput. Sci. Appl., 12: 539-545.
48. Biswas, N., K.M. Mohi Uddin, S.T. Rikta and S.K. Dey, 2022. A comparative analysis of machine learning classifiers for stroke prediction: A predictive analytics approach. Healthcare Anal., Vol. 2. 10.1016/j.health.2022.100116.
49. Wang, W., A.G. Rudd, Y. Wang, V. Curcin, C.D. Wolfe, N. Peek and B. Bray, 2022. Risk prediction of 30-day mortality after stroke using machine learning: A nationwide registry-based cohort study. BMC Neurol., Vol. 22. 10.1186/s12883-022-02722-1.
50. Wagner, M.J., C. Hennessy, A. Beeghly, B. French and D.P. Shah *et al*., 2022. Demographics, outcomes, and risk factors for patients with sarcoma and COVID-19: A CCC19-registry based retrospective cohort study. Cancers, Vol. 14. 10.3390/cancers14174334.
51. Shafiul Azam, M., M. Habibullah and H.K. Rana, 2020. Performance analysis of various machine learning approaches in stroke prediction. Int. J. Comput. Appl., 175: 11-15.
52. Jamthikar, A., D. Gupta, N.N. Khanna, L. Saba, J.R. Laird and J.S. Suri, 2020. Cardiovascular/stroke risk prevention: A new machine learning framework integrating carotid ultrasound image-based phenotypes and its harmonics with conventional risk factors. Indian Heart J., 72: 258-264.
53. Chen, J., Y. Chen, J. Li, J. Wang, Z. Lin and A.K. Nandi, 2022. Stroke risk prediction with hybrid deep transfer learning framework. IEEE J. Biomed. Health Inf., 26: 411-422.
54. Wang, K., T. Hong, W. Liu, C. Xu and C. Yin *et al*., 2023. Development and validation of a machine learning-based prognostic risk stratification model for acute ischemic stroke. Sci. Rep., Vol. 13. 10.1038/s41598-023-40411-2.
55. Zhou, J., S. Lee, Y. Liu, J.S.K. Chan and G. Li *et al*., 2023. Predicting stroke and mortality in mitral regurgitation: A machine learning approach. Curr. Probl. Cardiol., Vol. 48. 10.1016/j.cpcardiol.2022.101464.
56. Gomes, B., M. Pilz, C. Reich, F. Leuschner, M. Konstandin, H.A. Katus and B. Meder, 2021. Machine learning-based risk prediction of intrahospital clinical outcomes in patients undergoing TAVI. Clin. Res. Cardiol., 110: 343-356.
57. Tazin, T., M.N. Alam, N.N. Dola, M.S. Bari, S. Bourouis and M.M. Khan, 2021. Stroke disease detection and prediction using robust learning approaches. J. Healthcare Eng., Vol. 2021. 10.1155/2021/7633381.
58. Mostafa, S.A., D.S. Elzanfaly and A.E. Yakoub, 2022. A machine learning ensemble classifier for prediction of brain strokes. Int. J. Adv. Comput. Sci. Appl., 13: 258-266.
59. Fang, G., W. Liu and L. Wang, 2020. A machine learning approach to select features important to stroke prognosis. Comput. Biol. Chem., Vol. 88. 10.1016/j.compbiolchem.2020.107316.
60. Pathanjali, C., G. Monisha, T. Priya, S.K. Ruchita and S. Bhaska, 2020. Machine learning for predicting ischemic stroke. Int. J. Eng. Res. Technol., 9: 1334-1337.
61. Telu, V.S. and V. Padimi, 2022. Optimizing predictions of brain stroke using machine learning. J. Neutrosophic Fuzzy Syst., 2: 31-43.
62. Liu, J., Y. Sun, J. Ma, J. Tu and Y. Deng *et al*., 2021. Analysis of main risk factors causing stroke in Shanxi Province based on machine learning models. Inf. Med. Unlocked, Vol. 26. 10.1016/j.imu.2021.100712.
63. Zhang, Y., W. Song, S. Li, L. Fu and S. Li, 2018. Risk detection of stroke using a feature selection and classification method. IEEE Access, 6: 31899-31907.
64. Keerthy, K., 2022. Application of machine learning to analyze the risk factors of stroke. Int. J. Math. Appl., 10: 101-109.
65. Wu, Y. and Y. Fang, 2020. Stroke prediction with machine learning methods among older Chinese. Int. J. Environ. Res. Public Health, Vol. 17. 10.3390/ijerph17061828.
66. Macaulay, B.O., B.S. Aribisala, S.A. Akande, B.A. Akinnuwesi and O.A. Olabanjo, 2021. Breast cancer risk prediction in African women using random forest classifier. Cancer Treat. Res. Commun., Vol. 28. 10.1016/j.ctarc.2021.100396.
67. Fernandez-Lozano, C., P. Hervella, V. Mato-Abad, M. Rodríguez-Yáñez and S. Suárez-Garaboa *et al*., 2021. Random forest-based prediction of stroke outcome. Sci. Rep., Vol. 11. 10.1038/s41598-021-89434-7.

68. Qi, Y., 2012. Random Forest for Bioinformatics. In: Ensemble Machine Learning: Methods and Applications, Zhang, C. and Y. Ma (Eds.), Springer, New York, ISBN: 978-1-4419-9326-7, pp: 307-323.
69. Shobayo, O., O. Zachariah, M.O. Odusami and B. Ogunleye, 2023. Prediction of stroke disease with demographic and behavioural data using random forest algorithm. Analytics, 2: 604-617.
70. Saidul Islam, M., I. Hussain, M. Mezbaur Rahman, S.J. Park and M.A. Hossain, 2022. Explainable artificial intelligence model for stroke prediction using EEG signal. Sensors, Vol. 22. 10.3390/s22249859.
71. Kokkotis, C., G. Giarmatzis, E. Giannakou, S. Moustakidis and T. Tsatalas *et al*., 2022. An explainable machine learning pipeline for stroke prediction on imbalanced data. Diagnostics, Vol. 12. 10.3390/diagnostics12102392.
72. Holzinger, A., G. Langs, H. Denk, K. Zatloukal and H. Müller, 2019. Causability and explainability of artificial intelligence in medicine. WIREs Data Min. Knowl. Discovery, Vol. 9. 10.1002/widm.1312.