International Journal of Medical Informatics 203 (2025) 106041

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf

Review article

Machine learning techniques for stroke prediction: A systematic review of algorithms, datasets, and regional gaps

Afeez Adekunle Soladoye [a], Nicholas Aderinto [b], Mayowa Racheal Popoola [c], Ibrahim A. Adeyanju [a], Ayokunle Osonuga [d,e], David B. Olawade [f,g,h,*]

[a] Department of Computer Engineering, Federal University Oye-Ekiti, Ekiti, Nigeria
[b] Department of Medicine, Ladoke Akintola University of Technology, Ogbomoso, Nigeria
[c] Stroke Unit, Mid and South Essex NHS Foundation Trust, Westcliff-On-Sea SS0 0RY Northern Ireland, United Kingdom
[d] Coltishall Medical Practice, NHS GP Surgery, Norfolk NR12 7HA, United Kingdom
[e] Department of Primary Care, University of East Anglia, Norwich, United Kingdom
[f] Department of Allied and Public Health, School of Health, Sport and Bioscience, University of East London, London, Northern Ireland, United Kingdom
[g] Department of Research and Innovation, Medway NHS Foundation Trust, Gillingham ME7 5NY Northern Ireland, United Kingdom
[h] Department of Public Health, York St John University, London, Northern Ireland, United Kingdom

A R T I C L E I N F O

Keywords:
Stroke prediction
Machine learning
Deep learning
Systematic review
Artificial intelligence
Clinical decision support

A B S T R A C T

Background: Stroke is a leading cause of mortality and disability worldwide, with approximately 15 million people suffering strokes annually. Machine learning (ML) techniques have emerged as powerful tools for stroke prediction, enabling early identification of risk factors through data-driven approaches. However, the clinical utility and performance characteristics of these approaches require systematic evaluation.
Objectives: To systematically review and analyze ML techniques used for stroke prediction, systematically synthesize performance metrics across different prediction targets and data sources, evaluate their clinical applicability, and identify research trends focusing on patient population characteristics and stroke prevalence patterns.
Methods: A systematic review was conducted following PRISMA guidelines. Five databases (Google Scholar, Lens, PubMed, ResearchGate, and Semantic Scholar) were searched for open-access publications on ML-based stroke prediction published between January 2013 and December 2024. Data were extracted on publication characteristics, datasets, ML methodologies, evaluation metrics, prediction targets (stroke occurrence vs. outcomes), data sources (EHR, imaging, biosignals), patient demographics, and stroke prevalence. Descriptive synthesis was performed due to substantial heterogeneity precluding quantitative meta-analysis.
Results: Fifty-eight studies were included, with peak publication output in 2021 (21 articles). Studies targeted three main prediction objectives: stroke occurrence prediction (n = 52, 62.7 %), stroke outcome prediction (n = 19, 22.9 %), and stroke type classification (n = 12, 14.4 %). Data sources included electronic health records (n = 48, 57.8 %), medical imaging (n = 21, 25.3 %), and biosignals (n = 14, 16.9 %). Systematic analysis revealed ensemble methods consistently achieved highest accuracies for stroke occurrence prediction (range: 90.4–97.8 %), while deep learning excelled in imaging-based applications. African populations, despite highest stroke mortality rates globally, were represented in fewer than 4 studies.
Conclusion: ML techniques show promising results for stroke prediction. However, significant gaps exist in representation of high-risk populations and real-world clinical validation. Future research should prioritize population-specific model development and clinical implementation frameworks.

1. Introduction

Stroke ranks among the leading causes of mortality and disability worldwide. According to the World Health Organization [1], approximately 15 million people suffer strokes annually, resulting in 5 million deaths and another 5 million individuals living with permanent

* Corresponding author.
E-mail address: d.olawade@uel.ac.uk (D.B. Olawade).

https://doi.org/10.1016/j.ijmedinf.2025.106041
Received 12 June 2025; Received in revised form 3 July 2025; Accepted 6 July 2025
Available online 9 July 2025

disabilities [2,3]. Stroke manifests in two primary forms: ischemic, caused by a blockage in blood vessels (accounting for 85 % of cases), and hemorrhagic, resulting from a ruptured vessel (15 % of cases) [4]. Transient ischemic attacks also pose significant risks if untreated [5]. In the United States alone, stroke occurs every 40 s, with over 765,000 cases recorded annually, of which two-thirds are first-time strokes [2,6]. Developing countries, particularly in Sub-Saharan Africa, face an even graver burden, with stroke contributing to 87 % of stroke-related deaths due to limited access to advanced medical infrastructure, diagnostic tools, and timely interventions [7].

The rapid advancement of artificial intelligence (AI), particularly machine learning (ML), has revolutionized healthcare by enabling data-driven approaches to disease prediction, diagnosis, and treatment planning [8–11]. ML techniques leverage electronic health records (EHRs), bio-signals, and imaging data to accurately identify patterns and predict stroke risk [12–14]. Traditional ML algorithms, such as Support Vector Machines (SVM), Decision Trees (DT), and Random Forests (RF), have been widely used for structured data analysis, while deep learning models, including Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), excel in processing complex, unstructured data like brain MRIs and time-series signals [15,16]. These methods enable early identification of risk factors, facilitating proactive interventions that can mitigate stroke's devastating consequences. The availability of open-access datasets, such as those hosted on Kaggle, and hospital-based EHRs has further accelerated the development of ML-based prediction models, making them a cornerstone of modern medical data mining [17–22].

Despite the proliferation of ML applications in stroke prediction, the field lacks comprehensive reviews that synthesize the diverse methodologies, datasets, and performance metrics employed across studies. Existing studies often focus on developing novel ML models or evaluating specific algorithms, leaving critical gaps in understanding the broader landscape of stroke prediction techniques. For instance, while numerous studies have explored traditional ML algorithms like Logistic Regression and K-Nearest Neighbors, recent advancements in deep learning, such as Long Short-Term Memory (LSTM) networks, have gained traction for handling temporal and imaging data [15,16]. However, the comparative effectiveness of these approaches remains underexplored, particularly in terms of their predictive accuracy, generalizability, and applicability to diverse populations.

A significant concern is the geographical disparity in research output. This review identified fewer than four publications originating from Africa, despite Sub-Saharan Africa's disproportionately high stroke mortality rate [7]. This underrepresentation is likely due to limited access to EHR systems, insufficient research funding, and a lack of localized datasets, which hinders the development of Afrocentric ML models. The reliance on open-access datasets, such as those from Kaggle, raises additional concerns about data quality, representativeness, and potential biases, as the origins of these datasets are often unclear [17–22]. A systematic review is essential to map the global research landscape, evaluate the strengths and limitations of ML techniques, and identify strategies to address regional disparities. Such an analysis can guide researchers and policymakers in developing context-specific prediction models, particularly for high-risk regions like Sub-Saharan Africa, where early stroke detection could significantly reduce morbidity and mortality.

Moreover, integrating ML models into clinical practice remains a challenge, with issues such as data privacy, algorithmic bias, and interoperability with existing healthcare systems requiring further exploration. A systematic review can highlight best practices in data preprocessing, feature selection, and evaluation metrics, providing a foundation for improving model reliability and clinical utility [23,24]. This systematic review aims to comprehensively analyze machine learning techniques used for stroke prediction, evaluate their performance and methodological approaches, and identify research trends and geographical disparities in the field to guide future research and clinical implementation, with specific objectives to: (1) systematically identify and analyze ML techniques employed for stroke prediction in peer-reviewed literature published between 2013 and 2024, (2) evaluate the performance metrics and accuracy of different ML algorithms across various studies, (3) examine the characteristics and sources of datasets used in stroke prediction research, (4) assess data preprocessing techniques and feature selection methods employed, (5) investigate geographical distribution and publication trends in stroke prediction research, and (6) identify research gaps and provide recommendations for future studies, particularly for addressing regional disparities and improving clinical translation. This review specifically focuses on machine learning techniques for predicting the presence of stroke (risk assessment) using commonly available clinical data sources including electronic health records, medical imaging modalities (CT, MRI), and biosignals, rather than post-stroke treatment or rehabilitation applications.

2. Methods

This review evaluated ML techniques for stroke prediction. To ensure methodological rigor and transparency, it followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines.

2.1. Search strategy

A literature search was performed to identify open-access publications on ML-based stroke prediction published between January 1, 2013, and December 31, 2024. Five databases were searched: Google Scholar, Lens, PubMed, ResearchGate, and Semantic Scholar. These databases were selected for their comprehensive medical, scientific, and engineering literature coverage. PubMed and Lens provide access to peer-reviewed medical journals, ResearchGate and Google Scholar offer broad scholarly content, and Semantic Scholar specializes in scientific publications.

The search strategy employed a combination of keywords and Boolean operators to ensure relevance and specificity. The primary search terms included: "stroke prediction," "prediction of stroke," "prediction of ischemic stroke," "prediction of hemorrhagic stroke," "machine learning," "deep learning," "medical data mining," "medical informatics," "artificial intelligence," and "decision support systems." These terms were combined using Boolean operators as follows: (("prediction" AND "stroke") OR ("stroke" AND "machine" AND "learning") OR ("stroke" AND "prediction" AND "model") OR ("data" AND "mining")). The search was conducted between October 1, 2024, and December 15, 2024, to ensure recency.

2.2. Study selection

2.2.1. Eligibility criteria

Studies were included if they: (1) focused on stroke prediction using ML or deep learning techniques, (2) were published in peer-reviewed journals between 2013 and 2024, (3) were open-access, (4) included an abstract, and (5) were written in English. Exclusion criteria included: (1) studies not involving ML or deep learning, (2) non-journal publications (e.g., conference papers, books, or grey literature), (3) non-open-access articles, (4) studies lacking abstracts, and (5) studies focused on stroke treatment or rehabilitation rather than prediction.

2.2.2. Selection process

The selection process followed a multi-stage approach, as outlined in Fig. 1 (PRISMA flowchart). Initial searches across the five databases yielded the following results: Google Scholar (272 articles), Lens (54 articles with filters: open-access, has abstract, cited by scholarly works; reduced to 186 with journal article filter), PubMed (97 articles using "stroke prediction"), ResearchGate (75 articles), and Semantic Scholar

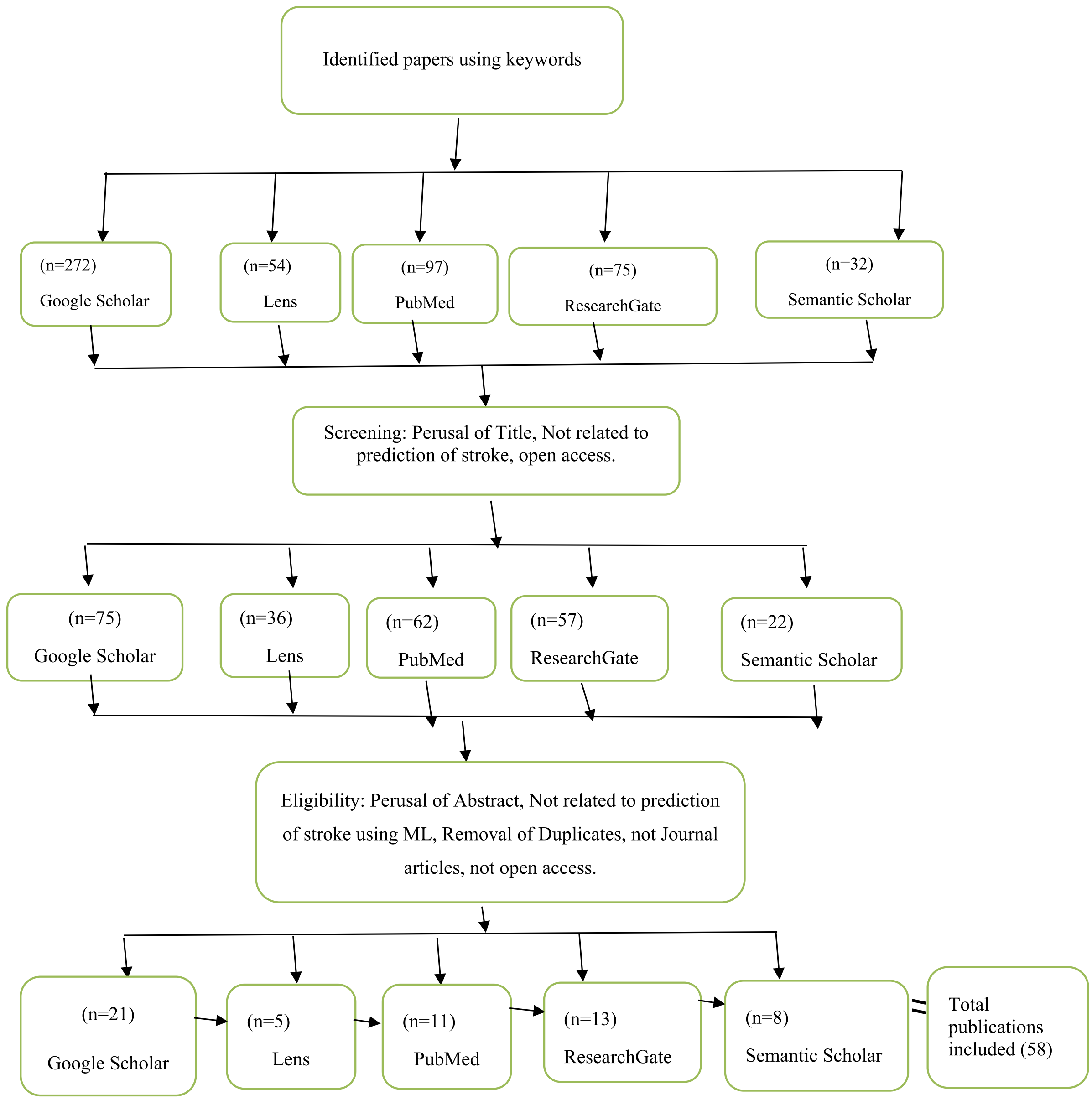

Fig. 1. Flow chart of the systematic review of relevant publications on the prediction of stroke with Machine Learning Approaches.

(32 articles). Abstracts were screened by two independent reviewers to assess eligibility based on the inclusion criteria. Discrepancies were resolved through discussion or consultation with a third reviewer. Full texts of potentially eligible articles were retrieved and further evaluated. Database duplicates were identified and removed using reference management software (Zotero). After screening and full-text review, 58 articles were included for analysis, comprising 21 from Google Scholar, 5 from Lens, 11 from PubMed, 13 from ResearchGate, 8 from Semantic Scholar, and four from additional Google searches.

2.3. Data extraction

Data were extracted using a standardized template to ensure consistency. Extracted variables included: (1) publication details (author, year, journal, country of affiliation), (2) dataset characteristics (source, size, type [structured, bio-signal, imaging], accessibility [open-access, self-acquired]), (3) ML methodologies (preprocessing techniques, feature selection/extraction methods, classification algorithms), (4) evaluation metrics (accuracy, sensitivity, specificity, F1 score, AUC), and (5) key findings (algorithm performance, limitations). One reviewer performed data extraction and was verified by a second to minimize errors. Discrepancies were resolved through consensus.

2.4. Bibliometric analysis

A bibliometric analysis was conducted to map publication trends and geographical distributions. Metrics included publication year, author affiliation country, and journal source. The study identified temporal trends (e.g., peak publication years) and geographical representation (e.g., number of publications by country). Data were visualized using bar charts and tables. Software tools, including R (version 4.4.1) and Excel, were used for quantitative analysis and visualization.

2.5. Data synthesis

Data were synthesized narratively due to heterogeneity in datasets, algorithms, and evaluation metrics, precluding *meta*-analysis.

3. Results

3.1. Publication Trends, Sources, and characteristics

Among the 58 included studies, 2021 had the highest publication output with 21 articles, reflecting a peak in ML and deep learning research for stroke prediction. This was followed by 2020 (9 articles), 2019 (7 articles), and 2022 (6 articles), with a decline in publication frequency from 2017 and earlier. Geographically, Asian countries dominated the research landscape, with India leading with 12 publications, followed by China (10) and South Korea (6). The United States contributed four publications, while Bangladesh and Australia had 3. Countries with 2 publications included Saudi Arabia, the United Kingdom, Ireland, Iran, Japan, and Portugal, whereas Thailand, Mexico, Malaysia, Morocco, and Egypt each had 1. This distribution is visualized in Fig. 2, which excludes countries with single publications for brevity.

The 58 studies were published across 22 peer-reviewed journals. Prominent journals included IEEE Journal of Biomedical and Health Informatics, JMIR Medical Informatics, Journal of the American Heart Association, Journal of the American Medical Informatics Association, Journal of Healthcare Engineering, International Journal of Advanced Computer Science and Applications, Frontiers in Genetics, and Frontiers in Aging Neuroscience. The journal "Sensors" published the highest number of individual articles (3). This variety of publication venues reflects the interdisciplinary nature of ML-based stroke prediction, spanning medical, informatics, and engineering fields.

3.2. Typical machine learning methodology for prediction of stroke

The use of ML to predict stroke can typically be categorized into four main stages: (1) Dataset Collection, (2) Data Preprocessing, (3) Classification, and (4) Model Evaluation, as illustrated in Fig. 3.

3.2.1. Dataset acquisition

Data acquisition is a critical first step in any ML task, as model performance depends heavily on the quality and availability of the input data. In the medical field, the advancement of EMRs and medical data mining has facilitated easier access to patient data for research.

3.2.1.1. Self-acquired datasets. Due to the sensitive nature of medical information, many datasets are collected directly from patients by trained professionals. These datasets are typically stored in hospital records and made available to researchers under strict confidentiality agreements. Such data are not publicly accessible and must be deleted after use by secondary researchers. Several stroke-related studies have used primary data from hospitals [25–29], while others have obtained secondary data under formal agreements [30–33].

3.2.1.2. Open-access datasets. Open-access datasets, such as those hosted on Kaggle, offer freely available EMRs for research and model development. Kaggle, a Google LLC subsidiary, serves as a hub for data scientists and ML practitioners, offering diverse datasets, including stroke data used in various studies [17–22]. Additional public sources include the Shenzhen Health Information Big Data Platform [34], the Cleveland database [35], WHO datasets [36], the International Stroke Trial (IST) dataset [24,37], and the Korean National Health Insurance Service (KNHIS) dataset [38]. Some datasets consist of structured clinical data, while others involve biosignal-based data such as EEG, ECG, EMG, and motion data [38]. Deep learning studies, in particular, often utilize medical imaging data, typically self-collected, from MRI or CT scans. For instance, one study reports using brain MRI scans acquired via a Siemens Skyra scanner [16].

Table 1 provides a comprehensive summary of the available datasets used for stroke prediction across the reviewed studies.

3.2.2. Data preprocessing techniques

Due to medical datasets' large volume, heterogeneity, and complexity, preprocessing is essential before applying machine learning algorithms. Preprocessing aims to convert raw data into a clean, integrated, and structured format suitable for analysis. This typically involves four key steps: data integration, cleaning, transformation, and reduction. The reviewed literature primarily addressed data integration and cleaning, as illustrated in Fig. 4.

3.2.2.1. Data integration. Medical datasets often originate from multiple sources, especially in low-resource settings where individual hospitals may not possess sufficient data. Data integration combines datasets from various institutions to ensure robust model training. However, sharing clinical data involves stringent privacy regulations, making direct access challenging. Consequently, some studies applied transfer learning to utilize correlated data from different hospitals, such as integrating EHRs from three hospitals in China for stroke prediction [39,40].

3.2.2.2. Data cleaning. Raw medical data frequently contains

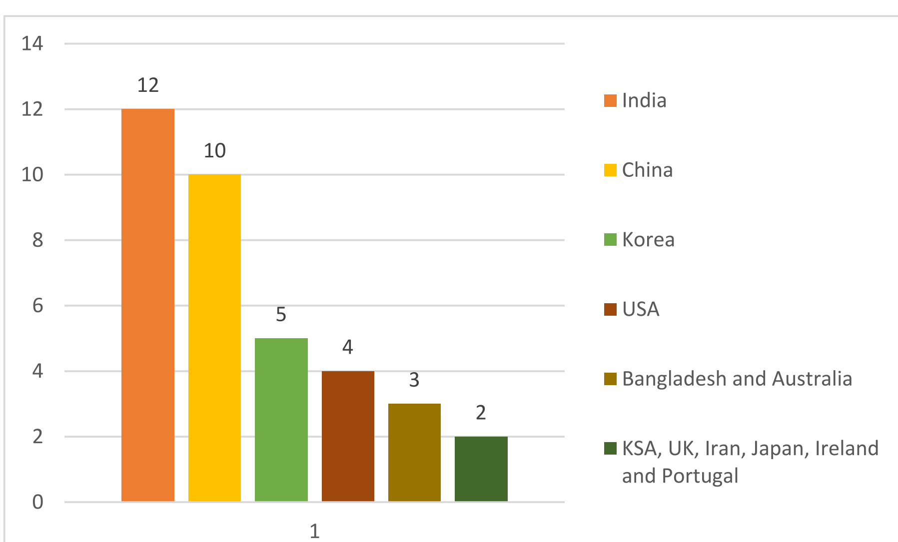

Fig. 2. Geographical distribution of the selected publications.

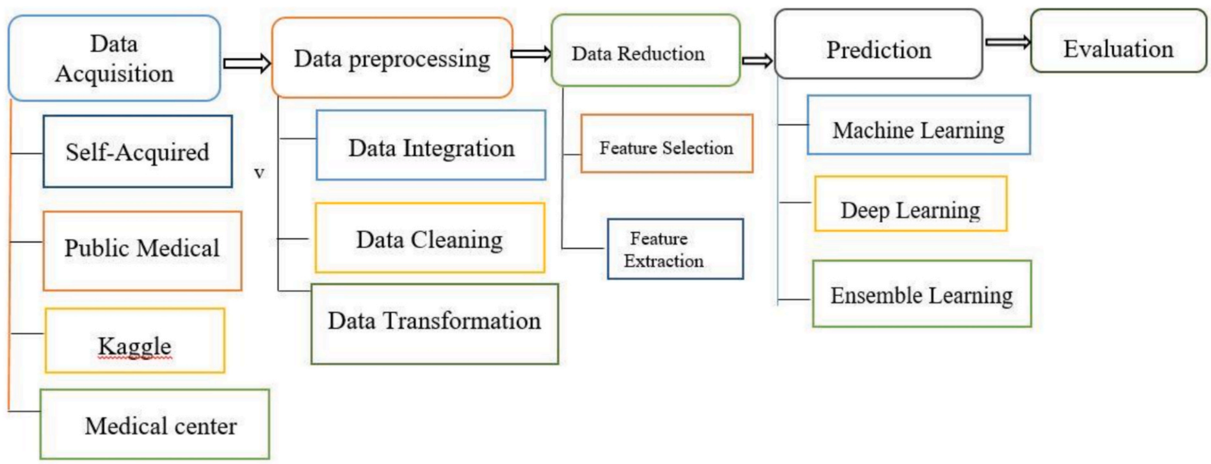

Fig. 3. Architecture of stroke prediction.

Table 1
Summary of the available dataset used for prediction of stroke.

S/N	Author	Dataset	Description
1	[33]	Cardiovascular Health Study	796 features and 4, 988 patients with 299 occurrences of stroke and others are healthy
2	[32]	Collected from King Abdulaziz Medical City (KAMC)	Dataset of patients diagnosed in 2016, The data set consists of 969 patients, 69 S free while 899 are stroke patients. The data set contains 360 females (33 S free and 327 S cases), 607 males (36 S free and 571 S cases) with 1004 attributes but 147 attributes were later used.
3	[19]	Healthcare Dataset Stroke Data uploaded by Saumya Agarwal available on Kaggle	This dataset contains data of 43,400 patients with 11 different attributes
4	[37]	International Stroke Trial (IST) available on IST website	19,435 patients from 467 hospitals in 36 countries
5	[17]	EHR released by McKinsey & Company for their healthcare hackathon challenge available on Kaggle	The dataset contains medical records of 29,072 patients with a total of 11 attributes
6	[24]	Third International Stroke Trial (IST-3)	It contains a total of 266 variables of 3035 patients older than 18 years where 1617 are above 80 years
8	[34]	Stroke data from Shenzhen Health Information Big Data Platform	This health platform have access to over 4000 health institutions, including 85 hospitals more than 650 community health service centers. It contains a total of 204,687 patients, 21,493 S and 183,194 non stroke patients.
9	[,,22,23,36]	"Stroke prediction dataset", "WHO stroke dataset" available on Kaggle	Contains records of 5110 patients with 11 attributes and outcome
10	[38]	Korean Health National Insurance (KHNIS) dataset	KHNIS database contains complete medical information of more than 50 million Koreans from which only the demographic and medical history of the patients were collected for building the model

inconsistencies, imbalances, outliers, and missing values that may impair model performance. Several preprocessing strategies were identified:

a. **Imbalanced Datasets:** A common challenge in stroke prediction is the class imbalance between stroke and non-stroke cases. Oversampling techniques such as SMOTE (Synthetic Minority Oversampling Technique) were used to generate synthetic minority samples [20–22]. Borderline-SMOTE, which focuses on minority samples near decision boundaries, was also employed [24,41]. In contrast, random down-sampling was used in some studies [17,26] to reduce the size of the majority class when oversampling was infeasible.
b. **Outliers and Noisy Data:** Outliers—data points that deviate significantly from others—can distort model predictions. Some studies removed or replaced them with null values [34], while others relied on algorithms like Artificial Neural Networks that are inherently robust to noise [17]. Clustering methods, such as X-means, were also applied to detect outliers using local density-based measures [42,43].
c. **Missing Values:** Missing data is prevalent in medical datasets due to incomplete records or errors during data collection. Imputation techniques such as mean, median, linear regression, and regularized expectation–maximization were used [17,23,33,35]. Random forest models, which can natively handle missing values, were also employed [34,43,44]. Alternatively, some studies excluded entries with missing values [37] or used zero-imputation for non-applicable attributes [45].

3.2.2.3. Data transformation. Data transformation involves converting raw data into a format suitable for machine learning models. This step enhances data compatibility with classification algorithms, minimizes errors, and improves prediction accuracy. Researchers applied Z-score standardization to large-scale features [24,42], while others used the StandardScaler method for dataset normalization [35].

3.2.2.4. Data discretization. Many classification algorithms require input features in a discrete format. Attributes such as gender, headache, and dizziness, which are often categorical or textual, are transformed using: Label Encoding – Assigns integer values to categorical variables [17,22,35,45]; One-Hot Encoding – Converts categories into binary vectors [24], allowing algorithms to process the data without assuming ordinal relationships.

3.2.3. Feature selection and extraction

Due to the vast size of medical datasets, feature reduction is essential

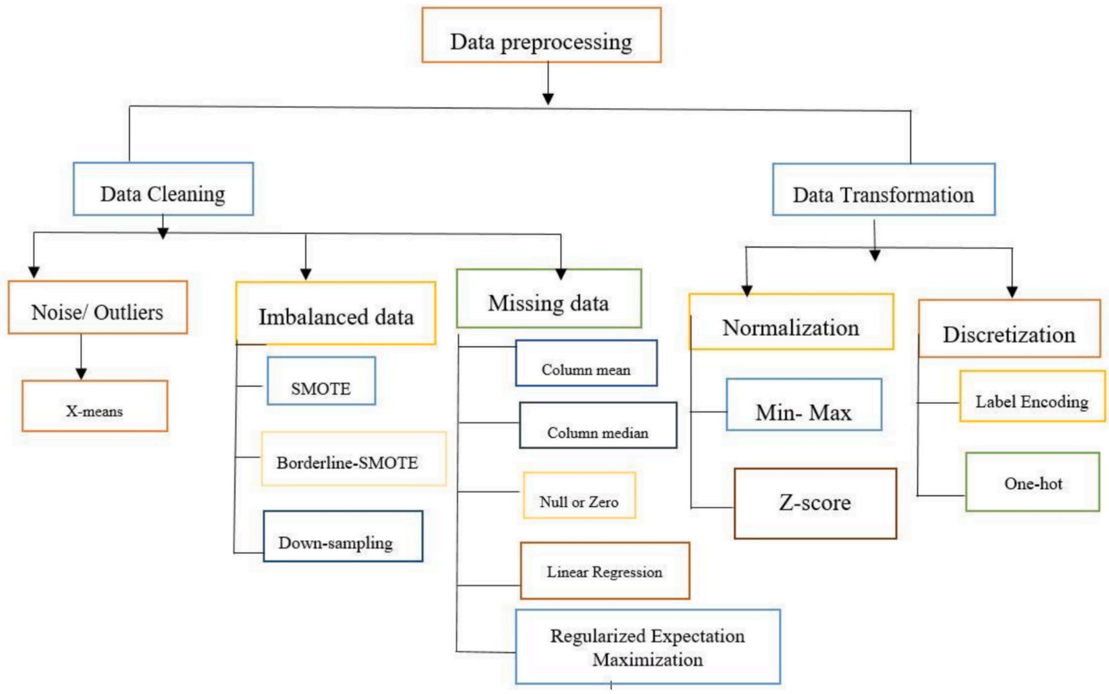

Fig. 4. Data preprocessing techniques.

to eliminate noise, reduce computational costs, and improve model learning. Feature selection aims to retain only the most relevant input variables, thus improving model performance and efficiency. Fig. 5 illustrates the various feature selection and extraction methods employed in the reviewed studies.

3.2.3.1. Feature selection methods. **Forward Feature Selection:** Starts with no features and incrementally adds one at a time, retaining it if it improves model performance. One study applied this with SVM but reported vulnerability to overfitting [33].

Backward Feature Selection (Backward Elimination): Begins with all features and removes one at a time, stopping when further removal reduces performance. This technique was used guided by p-values [23].

Pearson Correlation Coefficient: Measures the linear relationship between two variables and helps identify features significantly influencing an outcome. Studies found weak correlations among features, suggesting each variable had independent predictive value [17,37]. As a

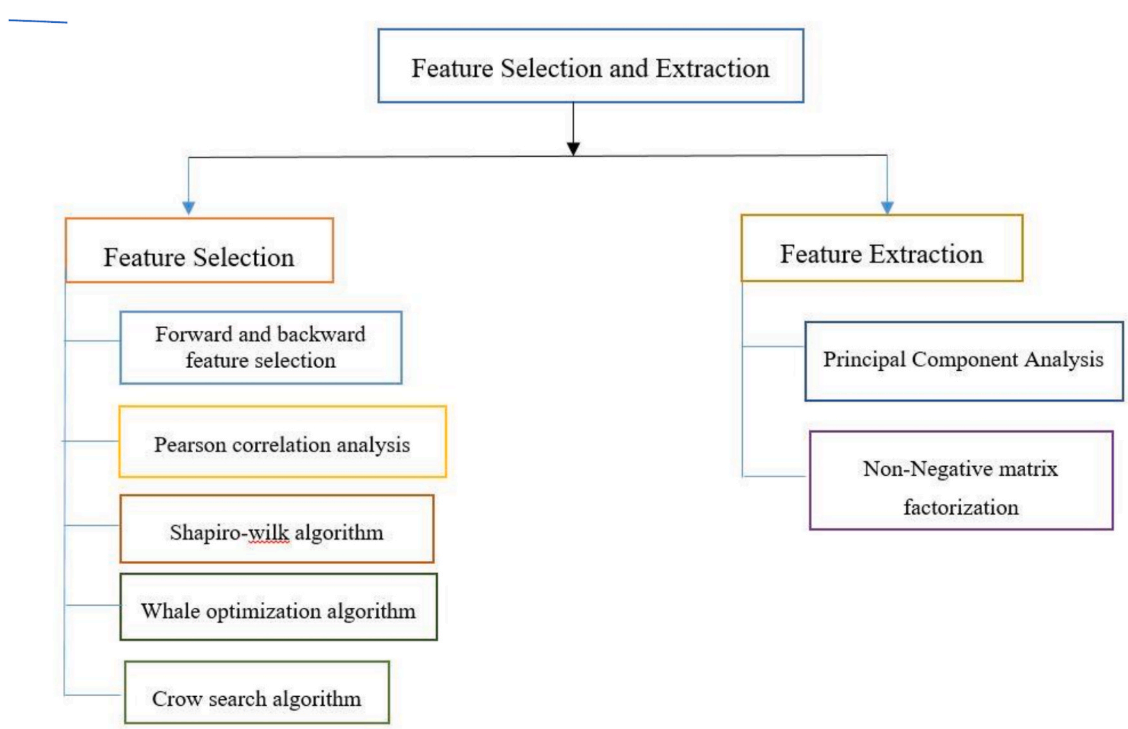

Fig. 5. Feature Selection and Extraction methods.

result, all 49 features were retained in both studies.

Shapiro-Wilk Test: Commonly used in medical research to assess the normality of data distributions and determine feature importance. With its high statistical power, it can identify features with minimal relevance. For instance, one study used this test and found only two features to be of limited value, ultimately retaining all features for model development [37].

Whale Optimization Algorithm (WOA): A metaheuristic inspired by the bubble-net hunting strategy of humpback whales. It mimics the search and encircling behavior of whales to optimize feature subsets. One study used WOA for feature extraction, leading to improved model performance [35].

Crow Search Algorithm (CSA): Based on the intelligent behavior of crows in hiding and retrieving food. It uses adaptive memory and strategic updates to eliminate irrelevant features. CSA was combined with WOA to refine features for stroke prediction, achieving enhanced accuracy [35].

3.2.3.2. Feature extraction. Feature extraction reduces dataset dimensionality by creating new, informative variables from original features. This process simplifies data, decreases computational load, and preserves critical patterns necessary for accurate predictions.

Principal Component Analysis (PCA): PCA transforms correlated features into a smaller set of uncorrelated components, capturing most of the variance in the original dataset. For instance, one study applied PCA to reduce 1004 features from patient records to 147, resulting in an accuracy improvement of over 8 % [32]. Similarly, another study grouped features into two principal components and selected four high-variance attributes: age, heart disease, hypertension, and average glucose level, as key stroke predictors [17]. PCA thus improves both model speed and accuracy by reducing redundancy in the data.

Non-Negative Matrix Factorization (NMF): NMF is an unsupervised technique tailored for non-negative datasets, often used in biomedical applications for its interpretability. Unlike PCA, it decomposes the original matrix into two non-negative matrices, preserving meaningful data structure. This enables effective extraction of lower-dimensional features while maintaining real-world relevance. Variants such as convex NMF, regularized NMF, and non-negative rank factorization offer further flexibility and control [46].

3.3. Machine learning classification algorithms employed in the prediction of stroke

Following data preprocessing in the disease prediction pipeline, the next step involves selecting an appropriate intelligent classifier to generate accurate predictions based on past patient records. Medical data mining, which integrates computing with healthcare, primarily relies on ML and its subset, DL. While ML enables systems to learn from data with minimal human intervention, DL uses neural networks to solve complex problems and excels in processing large datasets, including images, audio, and video. Traditional ML algorithms fall into two categories: supervised and unsupervised learning. Supervised learning involves training with labeled data (known outcomes) and is commonly used for classification tasks. In contrast, unsupervised learning identifies patterns or clusters in data without known outcomes. Based on reviewed literature, the most commonly used classifiers for stroke prediction include Artificial Neural Networks (ANN), Support Vector Machines (SVM), Random Forest (RF), K-Nearest Neighbors (KNN), Decision Trees (DT), Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and Ensemble Learning methods. Fig. 6 shows the distribution of machine learning algorithms employed by different publications.

3.3.1. Logistic regression (LR)

Logistic Regression is a supervised learning probabilistic classifier, also called logit or logistic model. This model is often used for predictive or classification analytics. It is used to predict the probability of existence of an event using dataset of existed independent variables. This is done by measuring the relation that occurs between the dependent and independent variables and measuring the possibility of occurrence of an event by fitting it into the logistic curve. This model can either be binary or multinomial, binary in the sense that the dependent variable is bipartite in nature while the independent variables are either nominal or continuous and when the dependent variables is more than two classes/categories multinomial logistic regression is used for such analysis [47].

Its applications are found mostly in comparison with other traditional ML models. This can be found in some literatures [17,22,27,48–52]. A prospective cohort study was performed on 512,726 people in China to predict stroke where Logistic regression was among the model employed [28]. Another study conducted using Korean National Health Insurance Service (KNHIS) dataset, where LR was used to identify features with significance to occurrence of stroke [38]. Another type of logistic regression known as personalized logistic regression (PLR) was employed to predict stroke [42].

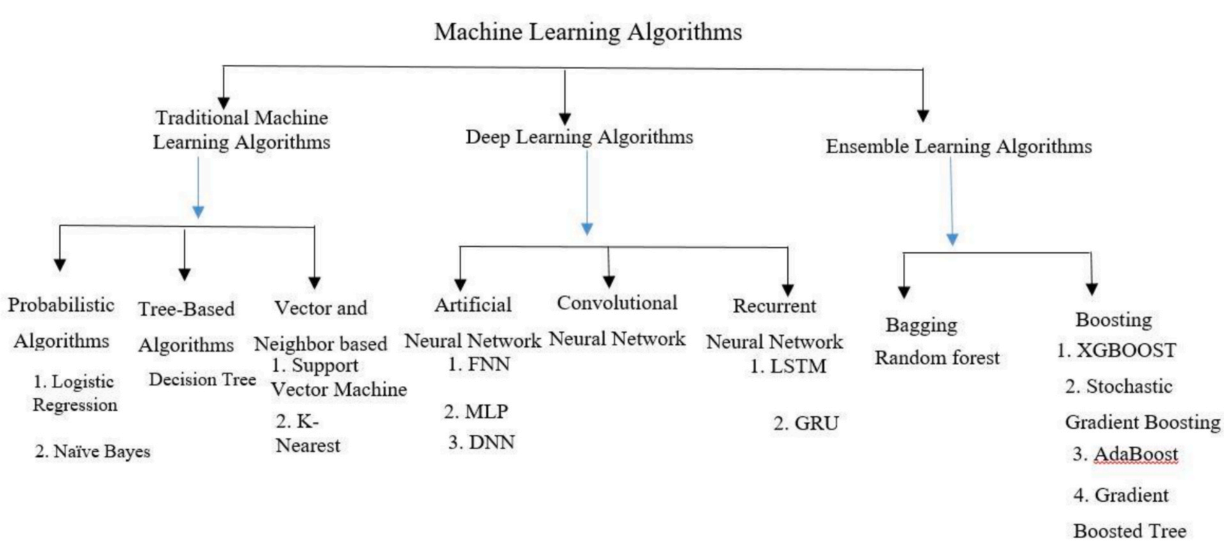

Fig. 6. Machine Learning Algorithms Employed by Different Publications.

Table 2 demonstrates the diversity and clinical potential of ML approaches in stroke prediction, with studies achieving consistently high performance metrics ranging from 84 % to 97 % accuracy across different clinical settings and data sources. Ensemble methods and deep learning architectures consistently outperformed traditional ML approaches, with notable successes including CNN-biLSTM for real-time EEG-based detection (96 % accuracy) and ensemble models achieving 97 % accuracy for population screening. Key predictive features consistently identified across studies include age, hypertension, glucose levels, and heart disease, suggesting these clinical markers have robust predictive value regardless of algorithm choice. However, the analysis reveals critical implementation challenges: most high-performing studies relied on small, region-specific datasets or open-access data of unclear provenance (particularly Kaggle datasets), while computationally intensive approaches like deep learning require specialized infrastructure that may limit deployment in resource-constrained settings.

3.3.2. Naïve Bayes (Bayesian) classifier

This is a probabilistic supervised learning algorithm like logistic regression for solving classification problems that predicts classes based on the probability of an object (predicts the probability of occurrence of an event based on the condition that another event has occurred). Bayesian Classifier as it is also known is based on Bayes theorem, majorly used for classification of text [45]. This algorithm is called Naïve because it presumes that occurrence of a phenotype is independent of occurrence of others like rise in blood pressure is independent of hypertension and this is the greatest drawback of this algorithm.

From the literatures, Naïve Bayes is usually used in the midst of other machine learning models to determine the one with the best performance among them, as it is seen in different literatures [26,45,49–51,53]. Gaussian Naïve Bayes was used in comparison with other traditional ML model to predict stroke [52].

3.3.3. Decision tree (DT)

One of the most important methods used for handling data with high dimensionality, this makes it widely used in the field of medical data mining, as medical records are usually of high dimension, and many researchers have used it for prediction of different life threatening diseases and it has been proven to be of good accuracy [53]. It is a supervised learning algorithm and a tree-like structured decision support tool that uses divide and conquer approach, which consists of nodes and leaves, decision tree nodes is gotten by comparing certain attributes with respect to a constant (record's attribute), in order to classify an unknown instance, it is traced down the tree based on the values of the attributes that have been successfully tested in the nodes, and when a leaf is reached during the routing, the instance will be classified based on the class assigned to that node [54].

Studies employed decision tree to predict stroke compared with other ML models [17,22–24,26,43,45,48,49,51]. C4.5 a well-known DT algorithm was used in comparison with other learning algorithms and it was implemented on Weka software [30,32].

3.3.4. K-Nearest neighbor (K-NN)

Like the two previously mentioned machine learning algorithms, KNN is also a supervised learning algorithm, lazy learner as it is widely known, this is because no model is learned and non-parametric algorithm, like DT it is also used for solving classification and regression problems. This algorithm makes its own prediction based on the proximity of the class label [53], as it assumes that similar and related things must exist in close proximity to each other. Euclidean distance is usually used to calculate the distance if these neighbors to each other. When a new instance is to be predicted, the closest to it among all the already

Table 2
Summary of key findings in ML-based stroke prediction.

Study	Algorithm(s)	Data Source	Key Predictors	Performance Metrics	Clinical Applications	Limitations
[17]	Random Forest, SVM	Kaggle EHR (5,000 records)	Age, hypertension, glucose level, BMI	Accuracy: 95 %, AUC: 0.92, F1: 0.90	Risk stratification in primary care using EHR data	Unclear dataset demographics; limited generalizability
[15]	CNN-biLSTM	EEG (self-acquired)	Biosignal power, relative values	Accuracy: 96 %, F1: 0.94, Sensitivity: 0.93	Real-time stroke detection in emergency settings	High computational cost; requires specialized EEG equipment
[35]	DNN with WOA, CSA	Hospital EHR (10,000 records, China)	Hypertension, heart disease, glucose	Accuracy: 97.34 %, AUC: 0.95	Risk prediction in hospital EHR systems; feature optimization	Small, region-specific dataset; validation needed
[52]	Ensemble (RF, SVM, KNN)	Kaggle EHR (5,000 records)	Age, glucose, BMI, smoking status	Accuracy: 97 %, MCC: 0.93, Specificity: 0.95	Enhanced accuracy for population-level screening	Potential overfitting; lacks external validation
[16]	CNN, LSTM	MRI (self-acquired, Siemens Skyra)	Radiology report vectors (word/sentence-level)	Accuracy: 90 %, AUC: 0.88	Prognosis prediction in radiology departments	Limited to MRI availability; complex preprocessing
[38]	DNN	KNHIS EHR (500,000 records, Korea)	Atrial fibrillation, age, clinical data	AUROC: 0.727, Sensitivity: 0.70	Risk prediction for AF patients in large-scale EHR systems	Outperformed traditional scores but limited to AF cohort
[28]	Logistic Regression, GBT	EHR (0.5 M Chinese adults)	Age, blood pressure, diabetes	Accuracy: 85 %, AUC: 0.80	Cohort-based risk assessment in public health	Large but region-specific dataset; moderate accuracy
[23]	SVM, RF, KNN	Kaggle EHR (4,000 records)	Hypertension, heart disease, smoking	Accuracy: 93 %, F1: 0.91	Primary care risk scoring; robust to noisy data	Unclear data origin; potential bias in open-access data
[37]	CNN, LSTM, SVM	Mixed (EHR, MRI)	Age, hypertension, imaging features	Accuracy: 94 %, AUC: 0.90	Hybrid model for risk and prognosis prediction	High computational requirements; data integration challenges
[34]	XGBoost	Shenzhen Health Platform (10,000 records)	Hypertension, glucose, cholesterol	Accuracy: 84.78 %, AUC: 0.922	Risk prediction for hypertensive patients	Limited to specific population; moderate accuracy
[31]	CNN, RNN	EHR (hospital, self-acquired)	Diagnoses, exam results	Accuracy: 84 %, Sensitivity: 0.82	EHR-based stroke prediction in hospital settings	Moderate performance; data privacy concerns
[32]	MLP, C4.5	EHR (self-acquired, 1,000 records)	Clinical data, demographics	Accuracy: 88 %, F1: 0.85	Small-scale hospital risk prediction	Small dataset; limited generalizability
[18]	Distributed ML (Spark)	EHR (self-acquired, 5,000 records)	Age, hypertension, heart disease	Accuracy: 92 %, AUC: 0.89	Scalable risk prediction for large datasets	Requires advanced infrastructure; data privacy issues
[22]	RF, LR, KNN	Kaggle EHR (5,000 records)	Age, glucose, smoking status	Accuracy: 94 %, AUC: 0.91	Cost-effective risk screening in primary care	Open-access data quality concerns
[45]	DT, KNN, Naïve Bayes	Kaggle EHR (3,000 records)	Hypertension, glucose, gender	Accuracy: 90 %, F1: 0.88	Simple models for resource-constrained settings	Lower accuracy; limited feature set

classified instances is calculated using Euclidean distance or Manhattan distance, the class of the instance that is closest to it is the solution to the new instance we're trying to predict. The k in the K-NN implies the number of the nearest neighbor to put into consideration in the voting process [55].

This classifier was used by many researchers for predicting stroke, such application can be found in literatures [23,30,45,48,49,51,52].

3.3.5. Support vector Machine (SVM)

This is one of the most common type of supervised learning with associated learning algorithms which is usually used for classification and solving regression problems, but mainly used and determined for classification. SVM is based on the derivation of hyperplane and support vector. In a dataset, there are usually many decision boundaries that separate different class labels, these labels are usually related based on their attributes, the best decision boundary that helped in accurate classification of this multiple decision boundaries is what is known as the hyperplane. There will be some points that have effect on the position of the hyperplanes and are also close to them on the plane, these data points are referred to as support vector. Many researchers have employed SVM for the prediction of stroke which can be found in literatures [21,23,24,37,44,48,49,52].

3.3.6. Artificial neural network (ANN)

It is sometimes shortened as neural network, ANN is a branch of artificial intelligence (AI) space, which is designed to replicate the functionality of human nervous system or in better term the biological neural network of the human brain [56], as a benchmark to develop a suitable machine learning algorithm that proffers solution to complex prediction and patterns problems. It is designed to consist of three layers input layer where the data are fed into the algorithm to perform its work, hidden layer which can be said to be distillation layer, where some important and relevant information are selected from the inputs, and the activation function applies weight to the input and send them over to the output layer and output layers is the layer which gives the outcome of the mined data. Different architectures of neural network used for stroke prediction are deep neural network (DNN), Multi-layer perceptron (MLP), backpropagation neural network and feed-forward neural network algorithms [21,23,44,53].

Multi-layer perceptron (MLP) was employed for prediction of stroke which was implemented in Weka software [32]. MLP was also used in another study [28]. ANN with feedforward and backpropagation neural network architecture was employed for prediction of stroke using an integrated dataset of medical screening and demographic data where a total of 147 S patients and 294 non-stroke individuals with a result accuracy of 84 % [31]. Similar study was found where hybrid optimization algorithm under deep neural network was employed for the prediction of cardiac stroke, whale optimization algorithm and Crow search algorithm were used for feature extraction/selection from the dataset, and the study yielded a good result with accuracy of 97.34 % [35].

Deep neural network was employed to predict the occurrence of ischemic stroke in Atrial fibrillation (AF) patients, the algorithm (AUROC $= 0.727 \pm 0.003$) was found to outperform CHA2DS2-VASc score (AUROC $= 0.651 \pm 0.007$) [38]. ANN with feedforward neural network used bioelectrical signal for early detection and prediction of stroke disease [29]. Another study hybridized deep neural network with transfer learning consisting of 4 components: Generative instance transfer, Network weight transfer, Bayesian optimization and Active instance transfer for stroke risk prediction, this showed a better accuracy than other stroke risk prediction methods with an accuracy of 0.757 [40].

3.3.7. Convolutional neural network (CNN)

This is the most established algorithm among the deep learning models, it is a class of the neural network which have found its application mostly in computer vision and image recognition, as it takes images as form of input, its application has spread out in different fields while medical field is not an exception as well [57]. CNN is designed to process data with grid pattern or structured array of data like images, which has its basis from the formation and arrangement of animal visual cortex [57,58], it comprises of building blocks like convolution layers, pooling layers, and fully connected layers [59]. The convolution layer is the basic part of the CNN which is in charge of feature extraction from the input data, which is the combination of both linear (Convolution operation) and non-linear operations (activation function). The pooling layer is the layer where the dimension of the features extracted by the convolution layer is reduced in order to reduce the size and number of the subsequent parameters that can be learned. The outputs of this layer are usually vectors of one-dimensional array; they are mapped by subset of fully-connected layers connected to the output network [57].

CNN was employed for feature extraction of both streaming and structured data where ReLU was used as the activation function [60]. CNN was combined with LSTM and CNN with bidirectional LSTM to predict stroke disease using EEG raw data, power values, and relative values, where the combination of CNN and bidirectional LSTM (CNN-biLSTM) is found to produce the best result [15]. CNN and RNN architectures were used in a model to predict future diagnosis of stroke using electronic health records involving diagnoses and exam result data [31]. CNN and multi-CNN were employed for prediction of stroke prognosis using MRI radiology report which have been preprocessed by dividing it into word and sentence level approach in order to have a document vector suitable as inputs into these models [16]. Many researchers employed CNN for prediction of stroke [37].

3.3.8. Recurrent neural network (RNN)

This is a type of artificial neural network usually used in speech processing and for solving natural language processing problems. RNN is used to handle time series and sequential data, which makes it more applicable when solving temporal problems like speech recognition and image captioning. Some medical data are temporal and their sequential reading must be processed with temporal based learner, which makes different architectures/variants of RNN to find their application in prediction of some illness like stroke. RNN differs from other traditional deep neural network because of their memory, which allows them to take information from the previous input to effect the current input and output.

There are different types of RNN: one input to one output, one input to many outputs, many input to one output and many input to many output. The most widely used RNN architectures are Long and Short Memory (LSTM), Bidirectional Long and Short Memory (BI-LSTM) and Gated Recurrent Unit (GRU), these variants of RNN were developed to solve the gradient vanishing problem encountered by simple RNN. LSTM and BI-LSTM were employed as deep learning model to prognosis stroke using MRI radiology report preprocessed into a document vector [16]. LSTM was employed to predict stroke using EHR where null values and anomalies are eliminated using JAVA programming and this model yielded an accuracy of 0.9998 [61]. LSTM was similarly used in another study [37].

LSTM comprises neural networks and many memory blocks which is known as Cell connected in a chain-like manner. A simple LSTM will structurally contain a cell, and three gates-input, forget and output gate-that manipulate the memory (cell) by controlling the flow of information in and out of the cell, while the cell stores information over time [15,62]. A very good variant of LSTM is bidirectional LSTM (biLSTM) which is a model for processing sequence or time variant data which comprises of two LSTM model taking inputs in opposite direction (forward and backward directions) but connected to the same output. This is much better than unidirectional LSTM because information can be retained for a longer period of time step.

Gated recurrent unit like LSTM was developed to solve the problem of gradient vanishing in conventional RNN, just that GRU have just three gates without any internal cells that keeps information like in the LSTM, the information needed to be stored are incorporated into GRU hidden state. GRU was said to outperform LSTM due to faster execution time, fewer gates and better performance accuracy [29,63].

3.3.9. Ensemble learning (EL)

This type of learning is based on hybridization or two or more other Classifiers or experts together for a better and improved predictive performance of the individual models. They are usually a combination of weak classifiers which becomes strong after their combination and produce a more accurate performance and improved prediction [64]. Ensemble learning have found its application in various fields and medical data mining is not an exception due to the importance of being able to give an optimal model with improved prediction accuracy over the base models, among the most used ensemble learning techniques/ algorithms are Random Forest (RF), Extreme gradient boosting (XGBoost), Stochastic Gradient boosting and Adaptive Boosting (AdaBoost).

Most used ensemble learning model is random forest as it can be used to proffer solution to both classification and regression problems and due to its ability to handle noisy and imbalanced dataset as most electronic health record are usually incomplete. It was employed by many researchers which can be found in literatures [17,22–24,37,43,45,54]. Two CNN models were combined together for feature extraction from streaming and structure data and prediction of stroke probability so as to produce an accurate prediction [60].

Random Forest was employed to identify which token is important in the word-level approach by extracting the feature importance of the text vectors [16]. Combined ANN and SVM comprising of two tiers were employed to predict stroke with the first tier using ANN with feed-forward neural network architecture for classification using data from SRP database and the second tier used the same multi-SVM for classification where normal brain MRI was used for training but brain MRIs with lesions removed were used for both testing and cross validation [46]. Stochastic gradient boosting was utilized in a study where Z-transformation was used for standardization and X-means was used to remove the outliers from the dataset for accurate prediction of stroke disease [42]. XGBoost was used to build a 3-year stroke prediction model for hypertensive patients which was implemented in Python3.6.5 kernel and produced an accuracy of 0.8478 and AUC of 0.9220, XGBoost was employed with some other classifiers [23,34]. Gradient boosted tree (GBT) was used for prediction of stroke risk among Chinese elders [28], while XGBoost was used for prediction with an accuracy of 96 % which outperformed other models compared [65]. Traditional ML models like SVM, KNN, LR and RF were stacked to ensure more accurate prediction where Random Forest was used as the meta learner, this gave the best accuracy of 97 % compared to other models used [52].

3.4. Evaluation methods and metrics

Evaluating ML models is essential to ensure they perform as expected when deployed. This stage is crucial for determining model reliability, user acceptance, and potential for commercialization. It involves various evaluation methods and performance metrics used by researchers to assess model effectiveness.

3.4.1. Cross validation

For any model development data is a very important part as the models learn through the data and the part of the dataset that the model learns from is known as training data, this is done so that the model can understand and learn the pattern of the model. After learning, the model has to be evaluated to know its performance, training data is not recommended to be used for evaluating the model so as to prevent overfitting, the part of the dataset used for evaluation is known as test data. This shows that every dataset is split to train and test data.

Cross validation is more preferable to hold out evaluation because in cross validation the model learns at every point of both training and testing data which ensures that all the classes are equally represented during testing. This method can be applied for evaluation on every dataset irrespective of its size be it small or large. This approach is used by many literatures as their evaluation method [17,18,26,28,32,34,37,38,42,43,45,46]. 5-fold cross validation was used [15], while 10-fold cross validation was employed [32].

3.4.2. Evaluation metrics

This is the approach of quantifying the performance of the model, this is mostly applicable to classification problems that is used with supervised learning approach whose expected class is known but not made known to the algorithm during the testing phase. Most used metrics are Accuracy, Sensitivity, Specificity, F1score, Area under Curve.

Accuracy: This measures the overall effectiveness of the developed system and it is measured in percentage.

Sensitivity/Recall: This is the ratio of the number of positive classes classified correctly to the total number of positive classes.

Specificity: This is the opposite of sensitivity which depicts the proportion of negative classes that are correctly classified as negative classes.

Precision: It depicts the number of true positive (positive classes) predicted that really belong to the positive class.

F1 score: This is the harmonic mean of recall and precision.

Most literatures used part or all these metrics to evaluate their models, this can be found in many literatures [17,18,23,30,32,33,35,37,38,42,43,45,53,60]. Some literatures used Area Under Curve (AUC) together with the above metrics for their evaluation [22,24,26,34,40], while others employed miss rate and fallout rate for their evaluation [17] and Matthews Correlation Coefficient (MCC) [52].

Table 3 provides a summary of the publications that used different evaluation metrics.

4. Discussion

This systematic review of 58 open-access studies from 2013 to 2024 provides an analysis of ML techniques for stroke prediction, highlighting their methodologies, performance, and global research trends. The bibliometric analysis revealed a surge in ML-based stroke prediction research, with 2021 marking the peak publication year (21 articles), followed by 2020 (9 articles) and 2019 (7 articles). This trend reflects growing global interest in leveraging AI for healthcare, driven by advancements in computational power and data availability [8,12].

Geographically, Asian countries, particularly India (12 publications) and China (10 publications), dominated the research landscape, likely due to robust investments in health informatics and large patient populations [17,23]. In contrast, the scarcity of African contributions, fewer than four publications, despite Sub-Saharan Africa's high stroke mortality rate (87 % of stroke-related deaths in developing countries) [7], highlights a critical research gap. This underrepresentation may stem from limited access to EHRs, insufficient research funding, and a lack of localized datasets, which hinder the development of Afrocentric ML models [7,39]. The absence of African studies is particularly concerning given the region's unique risk profiles, such as higher prevalence of

Table 3
Summary of the publications that used different evaluation metrics.

S/N	Evaluation Metrics	Author(s)
1	Accuracy, Precision, Recall, F1Score	[5,24,27,29–31,34,39,40,45,46,48]
2	Area Under Curve (AUC)	[22,35,36,38,43]
3	Miss Rate and Fallout Rate	[3]
4	Matthews Correlation Coefficient (MCC)	[60]

hypertension and limited healthcare infrastructure, which necessitate context-specific prediction models.

Dataset acquisition emerged as a pivotal factor influencing ML model performance. The review identified two primary data sources: self-acquired datasets from hospitals and open-access datasets from platforms like Kaggle, WHO, and the Shenzhen Health Information Big Data Platform [17–22,34–36]. Self-acquired datasets, often collected under strict confidentiality agreements, offer high specificity but are limited by accessibility and ethical constraints [25–33]. While widely used due to availability, open-access datasets raise concerns about data quality, representativeness, and potential biases, as their origins are often unclear [17–22]. For instance, Kaggle datasets may lack documentation on patient demographics or data collection protocols, compromising model generalizability [17,20]. Deep learning studies frequently utilize imaging data (e.g., MRI, CT scans), which require specialized preprocessing and high computational resources, limiting their feasibility in low-resource settings like Sub-Saharan Africa [16]. The reliance on non-local datasets underscores the need for region-specific data to address population-specific risk factors and improve model accuracy.

Preprocessing techniques were critical for addressing the complexity and heterogeneity of medical datasets. Common strategies included data cleaning (e.g., handling missing values via imputation or deletion), transformation (e.g., normalization, Z-score standardization), and reduction (e.g., feature selection, PCA) [24,32,33]. Imbalanced datasets, a prevalent issue in stroke prediction, were mitigated using over-sampling methods like SMOTE or down-sampling, with SMOTE being particularly effective in generating synthetic minority samples [20–22,24]. Feature selection methods, such as Pearson Correlation Coefficient, Whale Optimization Algorithm (WOA), and Crow Search Algorithm (CSA), enhanced model efficiency by identifying relevant predictors like age, hypertension, and glucose levels [17,35,37]. PCA and Non-Negative Matrix Factorization (NMF) further reduced dimensionality, improving computational efficiency and model performance [32,46]. However, the variability in preprocessing approaches across studies highlights the need for standardized protocols to ensure reproducibility and comparability.

The review identified a diverse range of ML algorithms, with traditional methods (e.g., Logistic Regression, SVM, Random Forest, Decision Trees, Naïve Bayes, KNN) and deep learning models (e.g., CNN, RNN, LSTM) being widely employed [15,16,23,53]. Random Forest and SVM were frequently used for their robustness to noisy and imbalanced datasets, achieving high accuracy in structured data analysis [23,24,45]. Deep learning models, particularly CNN and LSTM, excelled in processing unstructured data like EEG signals and MRI scans, with CNN-biLSTM combinations yielding superior results for temporal data [15,16]. Ensemble learning methods, such as XGBoost and Random Forest, demonstrated enhanced predictive performance by combining multiple weak classifiers, with some studies reporting accuracies exceeding 95 % [35,52,65]. However, the lack of a *meta*-analysis due to methodological heterogeneity precluded definitive comparisons of algorithm performance. Evaluation metrics (accuracy, sensitivity, specificity, F1 score, AUC) were consistently reported, with cross-validation (e.g., 5-fold, 10-fold) ensuring robust model assessment [15,32,37]. Notably, deep learning models often outperformed traditional algorithms in large-scale datasets but required significant computational resources, posing challenges for resource-constrained settings.

Integrating ML models into clinical practice remains a significant challenge. Issues such as data privacy, algorithmic bias, and interoperability with existing healthcare systems were underexplored in the reviewed studies [8,39]. For example, models trained on biased datasets (e.g., predominantly non-African populations) may perform poorly in diverse settings, exacerbating health disparities [7,17]. Data privacy concerns, particularly with self-acquired datasets, necessitate compliance with stringent regulations (e.g., GDPR, HIPAA), which were rarely addressed in the literature [25–33]. Furthermore, the computational complexity of deep learning models limits their deployment in low-resource hospitals, particularly in Sub-Saharan Africa, where EHR systems are often absent [7]. Ethical considerations, such as ensuring equitable access to ML-based diagnostics, are critical for global implementation but require stakeholder collaboration to resolve.

5. Limitations of the review

This systematic review has several important limitations that should be acknowledged. First, the focus on open-access publications may have excluded high-quality studies, potentially introducing publication bias and limiting the comprehensiveness of findings. Second, the substantial heterogeneity in study designs, outcome definitions, data sources, and performance measurement approaches precluded quantitative *meta*-analysis, limiting our ability to provide definitive comparative assessments of algorithm performance. Instead, we conducted descriptive synthesis which, while comprehensive, cannot provide the statistical rigor of pooled estimates.

Third, while our emphasis on geographical distribution of research provides important insights into global research disparities, we acknowledge that race, ethnicity, and population-specific stroke prevalence may be more clinically relevant parameters for understanding model performance and generalizability. However, many reviewed studies, particularly those using open-access datasets like Kaggle, lacked detailed demographic metadata including race and ethnicity information, limiting our ability to systematically analyze these factors. Future reviews should prioritize these demographic parameters as data quality and reporting standards improve.

Fourth, the underrepresentation of studies from high-burden regions, particularly Sub-Saharan Africa and Latin America, limits the external validity of findings for global health applications and may not reflect the true performance of ML models in these critical populations. Fifth, the variability in performance metric reporting and validation methodologies across studies made direct comparisons challenging and may have affected the reliability of observed performance patterns.

Sixth, while we have provided a comprehensive synthesis of ML techniques and their performance, the clinical utility of these findings may be limited by the lack of real-world validation studies and implementation frameworks. Most studies remained at the research stage without consideration of clinical workflow integration, regulatory requirements, or cost-effectiveness analyses necessary for practical deployment.

Seventh, the restriction to English-language publications may have excluded relevant research from non-Anglophone countries, potentially missing important regional insights and culturally specific approaches. Eighth, the classification of clinical implementation readiness was based on reported study characteristics rather than formal regulatory or clinical validation criteria, which may not accurately reflect true deployment readiness.

Finally, the rapid evolution of ML techniques means that some recent methodological advances may not have been fully captured within the review period. The absence of standardized reporting guidelines for ML research in healthcare also limited our ability to conduct more rigorous quality assessments of individual studies, particularly regarding data quality, model validation, and clinical applicability.

6. Conclusion

The rise of AI, particularly ML, has advanced medical practice, from enabling autonomous surgeries to developing predictive models for various diseases. Stroke, a leading cause of death worldwide, has attracted considerable research attention, especially in early prediction and risk assessment using medical data mining techniques. This review highlights the extensive use of machine learning in stroke prediction, showcasing a variety of approaches, tools, and algorithms applied in recent studies.

We examined the full pipeline of stroke prediction, including data acquisition, preprocessing (e.g., cleaning, transformation, and reduction), and the use of both traditional machine learning algorithms (e.g., SVM, Decision Trees, Random Forest, Logistic Regression, Naïve Bayes, and Neural Networks) and deep learning methods (e.g., CNNs and RNNs). As medical datasets grow in size and complexity, deep learning is increasingly favoured for its capacity to handle large-scale, high-dimensional data. Commonly used implementation platforms include Weka, MATLAB, and Python environments such as Anaconda. Challenges such as noisy or imbalanced datasets, missing values, and data incompleteness were also addressed, with researchers applying various strategies to mitigate their impact on model performance.

However, a significant gap exists in contributions from African researchers. Despite higher stroke mortality rates in Africa, fewer than four related publications from the continent were found during this review. Additionally, many studies rely on publicly available datasets from platforms like Kaggle, where data origin and quality are often unclear, potentially compromising accuracy and generalizability.

Future research should prioritise the use of real-world clinical data obtained directly from hospitals and patients. The lack of functional EHR systems in many African healthcare institutions is a major limitation that needs urgent attention. Furthermore, greater involvement from African researchers is crucial to developing context-specific models that address regional health challenges. Integrating machine learning models into clinical systems holds great promise, but more work is needed to translate research findings into practical, deployable tools.

CRediT authorship contribution statement

Afeez Adekunle Soladoye: Writing – review & editing, Writing – original draft, Validation, Methodology, Investigation, Conceptualization. **Nicholas Aderinto:** Writing – review & editing, Writing – original draft, Validation, Methodology, Investigation. **Mayowa Racheal Popoola:** Writing – review & editing, Validation, Methodology, Investigation. **Ibrahim A. Adeyanju:** Writing – review & editing, Methodology, Investigation. **Ayokunle Osonuga:** Writing – review & editing, Writing – original draft, Methodology, Investigation. **David B. Olawade:** Writing – review & editing, Writing – original draft, Validation, Methodology, Investigation, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] World Health Organization. Stroke, Cerebrovascular accident. 2019. Available from: https://www.who.int/news-room/fact-sheets/detail/stroke.
[2] C.W. Tsao, A.W. Aday, Z.I. Almarzooq, A. Alonso, A.Z. Beaton, M.S. Bittencourt, et al., Heart Disease and Stroke Statistics-2022 Update: a Report from the American Heart Association, Circulation 145 (8) (2022) e153–e639.
[3] World Health Organization. The top 10 causes of death. 2020. Available from: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death.
[4] R.V. Krishnamurthi, T. Ikeda, V.L. Feigin, Global, regional and country-specific burden of ischaemic stroke, intracerebral haemorrhage and subarachnoid haemorrhage: a systematic analysis of the global burden of disease study 2017, Neuroepidemiology 54 (2020) 171–179.
[5] R. Nial, What are the different types of strokes? Healthline (2018).
[6] Centers for Disease Control and Prevention. Stroke. 2022. Available from: https://www.cdc.gov/stroke/facts.htm.
[7] E. Bertha, O. Adesola, I. Emmanuel, E. Udeme, Stroke mortality and its predictor in a Nigerian teaching hospital, Africa Health Sci. 15 (1) (2015) 74–81.
[8] H.G. Koh, G. Tan, et al., Data mining applications in healthcare, J. Healthc. Inf. Manag. 19 (2) (2021) 65–72.
[9] I. Yoo, P. Alafaireet, M. Marinov, K. Pena-Hernandez, R. Gopidi, J.F. Chang, et al., Data mining in healthcare and biomedicine: a survey of the literature, J. Med. Syst. 36 (4) (2012) 2431–2448.
[10] G. Philipp, J. Grunz, P.J. Kuhl, A. Kosmala, T.A. Bley, B. Petritsch, et al., Performance testing of a novel deep learning algorithm for the detection of intracranial hemorrhage and first trial under clinical conditions, Neurosci. Inf. 1 (2021) 100007.
[11] R. AlSaad, Q. Malluhi, S. Boughorbel, PredictPTB: an interpretable preterm birth prediction model using attention-based recurrent neural networks, Biodata Min. 16 (5) (2022) 1–19.
[12] G. Sivapalan, K. Nundy, S. Dev, B. Cardiff, D. John, ANNet: a lightweight neural network for ECG anomaly detection in IoT edge sensors, IEEE Trans. Biomed. Circuits Syst. 14 (6) (2020) 1297–1308.
[13] N.K. Chowdhury, M.K. Kabir, M. Rahman, S.S. Islam, Machine learning for detecting COVID-19 from cough sounds: an ensemble-based MCDM method, Comput. Biol. Med. 145 (2022) 105508.
[14] T. Wang, C. Zhang, A. He, W. Dong, Overview of crow search algorithm, J. Phys. Conf. Ser. 2258 (2022) 012017.
[15] Y. Choi, S. Park, J. Jun, C. Pyo, K. Cho, H. Lee, et al., Deep learning-based stroke disease prediction system using real-time bio signal, Sensors 21 (13) (2021) 4269.
[16] T.S. Heo, Y.S. Kim, J.M. Choi, Y.S. Jeong, S.Y. Seo, J.H. Lee, et al., Prediction of stroke outcome using natural language processing-based machine learning of radiology report of Brain MRI, J. Personal. Med. 10 (4) (2020) 286.
[17] S. Dev, H. Wang, C.H. Nwosu, N. Jain, B. Veeravalli, D. John, A predictive analytics approach for stroke prediction using machine learning and neural networks, Healthcare Anal. 2 (2022) 100032.
[18] Ahmed H, Abd-el ghany SF, Youn EMG, Omran NF, Ali AA. Stroke Prediction using Distributed Machine Learning Based on Apache Spark. Int. J. Adv. Sci. Technol. 2019;28(15):89-97.
[19] Y. Choi, J.W. Choi, Stroke Prediction using Machine Learning based on Artificial Intelligence, Int. J. Adv. Trend. Comput. Sci. Eng.. 9 (5) (2020) 8916–8921.
[20] M. Islam, S. Akter, M. Rokunojjaman, J.H. Rony, A.l. Amin, S. Kar, Stroke Prediction Analysis using Machine Learning Classifiers and Feature Technique, Int. J. Electron. Commun. Syst. 1 (1) (2021) 17–22.
[21] F.H. Hassan, M.A. Omar, Recurrent Stroke Prediction using Machine Learning Algorithms with Clinical Public Datasets: an Empirical Performance Evaluation, Baghdad Sci. J. 18 (4) (2021) 1406–1412.
[22] T. Tazin, M.N. Alam, N.N. Dola, M.S. Bari, S. Bourouis, M.M. Khan, Stroke Disease Detection and Prediction using Robust Learning Approaches, J. Healthcare Eng. 2021 (2021) 7633381.
[23] J.A.T. Rodríguez, Stroke prediction through Data Science and Machine Learning Algorithms, Available from: ResearchGate (2021) https://www.researchgate.net/publication/352261064.
[24] Q. Qin, X. Zhou, Y. Jiang, Prognosis Prediction of Stroke based on Machine Learning and Explanation Model, Int. J. Comput. Commun. Control. 16 (2) (2021) 4108.
[25] Q. Zhang, Y. Xie, P. Ye, C. Pang, Acute ischaemic stroke prediction from physiological time series patterns, Austral. Med. J. 6 (5) (2013) 280–286.
[26] S. Thammaboosadee, T. Kansadub, Data mining model and application for stroke prediction: a combination of demographic and medical screening data approach, Interdiscipl. Res. Rev. 14 (4) (2019) 61–69.
[27] L. Zhao, H. Wang, Z. Yang, B. Jiang, H. Li, Y. Wang, Multimodal retinal imaging for detection of ischemic stroke, Front. Aging Neurosci. 13 (2021) 615813.
[28] M. Chun, R. Clarke, B.J. Cairns, D. Clifton, D. Bennett, Y. Chen, et al., Stroke risk prediction using machine learning: a prospective cohort study of 0.5 million Chinese adults, J. Am. Med. Inform. Assoc. 28 (8) (2021) 1719–1727.
[29] R. Islam, S. Debnath, T.I. Palash, Predictive Analysis for Risk of Stroke Using Machine LEarning Techniques, 2021 International Conference on Computer, COmmunication, Chemical, Mater. Electron. Eng. (IC4ME2), (2021) 1–4.
[30] L. Amini, R. Azarpazhooh, M.T. Farzadfar, S.A. Mousavi, F. Jazaieri, F. Khorvash, et al., Prediction and control of stroke by data mining, Int. J. Prev. Med. 4 (2013) S245–S249.
[31] D. Teoh, Towards stroke prediction using electronic health record, BMC Med. Inf. Decis. Making 18 (2018) 127.
[32] O. Almadani, R. Alshammari, Prediction of stroke using data mining classification techniques, Int. J. Adv. Comput. Sci. Appl. 9 (1) (2018) 457–460.
[33] A. Khosla, Y. Cao, C.C. Lin, H. Chiu, J. Hu, H.L.H. Lee, An integrated machine learning approach to stroke prediction, in: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2010, pp. 183–192.
[34] Y. Yang, J. Zheng, Z. Li, Y. Li, Y. Cai, Accurate prediction of stroke for hypertensive patients based on medical big data and machine learning algorithms: retrospective study, JMIR Med. Inform. 9 (11) (2021) e30277.
[35] N.K. Al-Shammari, A.A. Alzami, M. Albadarn, S.T. Ahmed, S.M. Basha, A. S. Alshammari, et al., cardiac stroke prediction framework using hybrid optimization algorithm under DNN, Eng., Technol. Appl. Sci. Res. 11 (4) (2021) 7436–7441.
[36] S.S. Aundhakar, A.G. Patil, Stroke prediction using machine learning methodology, Int. J. Res. Eng., Sci. Manage. 4 (11) (2021) 107–109.
[37] G. Fang, Z. Huang, Z. Wang, Predicting ischemic stroke outcome using deep learning approaches, Front. Genet. 12 (2022) 827522.
[38] S. Jung, M. Song, E. Lee, S. Bae, Y. Kim, L. Lee, et al., Predicting ischemic stroke in patients with atrial fibrillation using machine learning, Frontiers in Bioscience (landmark Edition). 27 (3) (2022) 82.
[39] Lee J, Sattigeri P, Wornell G. Learning new tricks from old dogs: Multi-source transfer learning from pre-trained networks. In Proceedings of 2019 International Conference of Neural Information Processing Systems. 2019:4372–4382.
[40] J. Chen, T. Chen, J. Li, J. Wang, Z. Lin, A.K. Nandi, Stroke risk prediction with hybrid deep transfer learning framework, IEEE J. Biomed. Health Inform. 26 (1) (2022) 411–422.
[41] Han H, Wang W, Mao B. Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. International Conference on Intelligent Computing. 2005, pp. 878-887.

[42] A.K. Arslan, C. Colak, M.E. Sarihan, Different medical data mining approaches based prediction of ischemic stroke, Comput. Methods Programs Biomed. 130 (2016) 87–92.

[43] C. Pathanjali, T. Priya, G. Monisha, S. Bhaskar, K.R. Sudarshan, Machine Learning for predicting Ischemic Stroke, Int. J. Eng. Res. Technol. 9 (5) (2020) 1334–1337.

[44] B.P. Deepak Kumar, S. Yellaram, S. Kothamasu, S. Puchakayala, Heart Stroke Prediction using Machine Learning, Int. J. Creat. Res. Thought. 9 (6) (2021) a328–a332.

[45] Shoily TI, Islam T, Jannat S, Tanna SA, Alif TM, Ema RR. Detection of Stroke Disease using Machine Learning Algorithms. 10th ICCCNT 2019, IIT Kanpur, India. 2019.

[46] Jerrin TP, Jeena RS. Two Tier Prediction of Stroke using Artificial Neural Networks and Support Vector Machines. arXiv:2003.08354v2 [cs.LG]. 2020.

[47] H. Park, An introduction to logistic Regression: from Basic Concepts to Interpretation with Particular attention to Nursing Domain, J. Korean Acad. Nursing. 43 (2) (2013) 154–164.

[48] K.V. Harshitha, P. Harshitha, G. Gunjan, P. Vaishak, K.B. Prajna, Stroke Prediction using Machine Learning Algorithms, Int. J. Innovat. Res. Eng. Manage. 8 (4) (2021) 146–151.

[49] T. Rakshit, A. Shrestha, Comparative analysis and implementation of heart stroke prediction using various machine learning techniques, Int. J. Eng. Res. Technol. 10 (6) (2021) 664–670.

[50] N. Biswas, K.M.M. Uddin, S.T. Rikta, S.K. Dey, A comparative analysis of machine learning classifiers for stroke prediction: a predictive analytics approach, Healthcare Anal. 2 (2022) 100116.

[51] R. Gurjar, H.K. Sahana, C. Neelambika, B.S. Sparsha, S. Ramys, Stroke Risk Prediction using Machine Learning Algorithms, Int. J. Scient. Res. Comput. Sci., Eng. Informat. Technol. 8 (4) (2022) 20–25.

[52] S.A. Mostafa, D.S. Elzanfaly, A.E. Yakoub, A machine learning ensemble classifier for prediction of brain strokes, Int. J. Adv. Comput. Sci. Appl. 12 (12) (2022) 650–657.

[53] A. Sudha, P. Gayathri, N. Jaisankar, Effective analysis and predictive model of stroke disease using classification methods, Int. J. Comput. Appl. 43 (14) (2012) 26–31.

[54] I.H. Witten, E. Frank, M.A. Hall, Data Mining Practical Machine Learning Tools and Techniques, Morgan Kaufmann Publishers USA, 2011.

[55] S.B. Kotsiantis, I.D. Zaharakis, P.E. Pintelas, Machine learning: a review of classification and combining techniques, Artif. Intell. Rev. 26 (3) (2006) 159–190.

[56] N. Salman, A. Lawi, S. Syarif, Artificial Neural Network Backpropagation with Particle Swarm Optimization for Crude Palm Oil Price Prediction, IOP Conference Series: J. Phys.: Conf. Series. 1114 (2018) 012088.

[57] R. Yamashita, M. Nishio, R.K. Gian Do, K. Togashi, Convolutional neural networks: an overview and application in radiology, Insight. Imag. 9 (2018) 611–629.

[58] K. Fukushima, Neocognitron: a self organizing neural network model for a mechanism of pattern recognition unaffected by shift in position, Biol. Cybern. 36 (1980) 193–202.

[59] S.S. Nisha, N.M. Meeral, Applications of deep learning in biomedical engineering, in: Handbook of Deep Learning in Biomedical Engineering, 2021, pp. 195–220.

[60] Y. Liu, B. Yin, Y. Cong, The probability of ischaemic stroke prediction with a multi-neural-network model, Sensors 20 (2020) 4995.

[61] Chantamit-o-pas P, Goyal M. Long Short-Term Memory Recurrent Neural Network for Stroke Prediction. ResearchGate. 2018. Available from: https://www.researchgate.net/publication/326255515.

[62] Kalita D. An overview of Long Short Term Memory (LSTM). Analytics Vidhya. 2022. Available from: https://www.analyticsvidhya.com/blog/2022/03/an-overview-on0long-short-term-memory-lstm/.

[63] S. Dutta, J.K. Mandal, T.H. Kim, S.K. Bandyopadhyay, Breast cancer prediction using stacked GRU-LSTM-BRNN, Appl. Comput. Syst. 25 (2) (2020) 163–171.

[64] X. Dong, Z. Yu, W. Cao, Y. Shi, Q. Ma, A survey on ensemble learning, Front. Comput. Sci. 14 (2) (2020) 241–258.

[65] A.A. Rahim, A. Sunyoto, M.R. Arief, Stroke Prediction using machine learning method with extreme gradient boosting algorithm, Matrik: Jurnal Manajemen, Teknik Informatika, Dan Rekayasa Komputer. 21 (3) (2022) 595–606.