Multimodal Machine Learning for Stroke Prognosis and Diagnosis: A Systematic Review

Saeed Shurrab, Alejandro Guerra-Manzanares, Amani Magid, Bartlomiej Piechowski-Jozwiak, S. Farokh Atashzar, *Senior Member, IEEE*, and Farah E. Shamout

(Review paper)

***Abstract*—Stroke is a life-threatening medical condition that could lead to mortality or significant sensorimotor deficits. Various machine learning techniques have been successfully used to detect and predict stroke-related outcomes. Considering the diversity in the type of clinical modalities involved during management of patients with stroke, such as medical images, bio-signals, and clinical data, multimodal machine learning has become increasingly popular. Thus, we conducted a systematic literature review to understand the current status of state-of-the-art multimodal machine learning methods for stroke prognosis and diagnosis. Following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines during literature search and selection, our results show that the most dominant techniques are related to the fusion paradigm, specifically early, joint and late fusion. We discuss opportunities to leverage other multimodal learning paradigms, such as multimodal translation and alignment, which are generally less explored. We also discuss the scale of datasets and types of modalities used to develop existing models, highlighting opportunities for the creation of more diverse multimodal datasets. Finally, we present ongoing challenges and provide a set of recommendations to drive the next generation of multimodal learning methods for improved prognosis and diagnosis of patients with stroke.**

***Index Terms*—Stroke, multimodal clinical data, machine learning, deep learning.**

Received 12 December 2023; revised 27 June 2024 and 30 July 2024; accepted 17 August 2024. Date of publication 22 August 2024; date of current version 7 November 2024. This work was supported by Tamkeen through NYU Abu Dhabi Research Enhancement Fund, in part by the NYUAD Center for AI & Robotics under Grant CG010, in part by the Center for Interacting Urban Networks under Grant CG001, in part by the Center for Cyber Security under Grant G1104, in part by ASPIRE, the Technology Program Management Pillar of Abu Dhabi's Advanced Technology Research Council (ATRC), through ASPIRE Precision Medicine Research Institute Abu Dhabi (ASPIREPMRIAD) under Award VRI-20-10. *(Saeed Shurrab and Alejandro Guerra-Manzanares contributed equally to this work.) (Corresponding author: Farah E. Shamout.)*

Saeed Shurrab, Alejandro Guerra-Manzanares, and Farah E. Shamout are with the Computer Engineering Division at New York University Abu Dhabi, Abu Dhabi 129188, UAE (e-mail: saeed.shurrab@nyu.edu; alejandro.guerra@nyu.edu; farah.shamout@nyu.edu).

Amani Magid is with the Science and Engineering library at New York University Abu Dhabi, Abu Dhabi 129188, UAE (e-mail: am6087@nyu.edu).

Bartlomiej Piechowski-Jozwiak is with the Neurological Institute at Cleveland Clinic Abu Dhabi, Abu Dhabi 112412, UAE (e-mail: neurobart@gmail.com).

S. Farokh Atashzar is with the Electrical and Computer Engineering, Mechanical and Aerospace Engineering at New York University, New York, NY 11201 USA (e-mail: f.atashzar@nyu.edu).

Digital Object Identifier 10.1109/JBHI.2024.3448238

I. Introduction

STROKE is the most common form of cerebrovascular disease. It ranks as the second leading cause of death globally and the most common cause of major cognitive and sensorimotor disability in adults [1]. Patients who survive their first stroke are highly expected to experience recurrence, which can range from days to years [2]. Moreover, stroke survivors are prone to functional loss and emotional side effects that negatively impact quality of life. The global stroke-related burden figures showed a considerable increase in terms of incidence (70%), prevalence (102%), mortality (43%), and disability-adjusted life years (143%) between 1990 and 2019 [3]. The majority of the global stroke burden concentrates in low- and middle-income countries [4], emphasizing the need for improved stroke management and prevention to mitigate the associated economic and societal impact.

Stroke diagnosis refers to the process of identifying if a patient has a stroke, based on the analysis of focal neurological signs and symptoms of presumed vascular origin [5]. It requires timely and accurate therapeutic pathways to reduce brain damage [6]. Typically, an accurate diagnosis within 60 minutes from stroke onset results in better outcomes [7]. Clinical tests such as the Face Arm Speech Test (FAST), the National Institutes of Health Stroke Scale (NIHSS), and the Cincinnati Pre-hospital Stroke Scale (CPSS) are adopted in emergency rooms to triage stroke patients [8]. Neuro-imaging techniques such as Magnetic Resonance Imaging (MRI) and Computed Tomography (CT) are most effectively used to confirm the diagnosis. Despite the availability of advanced imaging tools, stroke diagnosis is especially difficult in situations where symptoms go unnoticed [9], when a recent stroke gets confused with a previous stroke [10], or due to variability in clinical judgment [8].

Stroke prognosis refers to the process of predicting the short- and long-term likelihood of adverse outcomes amongst patients with stroke, depending on the type of stroke. Ischemic stroke occurs when a cerebral artery experiences an embolic or thrombotic occlusion, whereas hemorrhagic stroke occurs due to cerebral artery rupture. Transient ischemic attack is a mini stroke that lasts for less than a few minutes and does not cause any tissue damage [11]. The 90 days modified Rankin Scale (mRS) is a well-known stroke prognostic tool that measures the level of disability in stroke patients after 90 days post-discharge. It

TABLE I
OVERVIEW OF EXISTING LITERATURE REVIEWS RELATED TO MACHINE LEARNING (ML) AND DEEP LEARNING (DL) FOR STROKE

Ref.	Year	Scope
[15]	2022	DL for stroke detection, lesion segmentation, and outcome prediction using neuro-imaging data
[16]	2022	ML and DL for ischemic stroke classification and segmentation using MRI data
[17]	2022	DL for ischemic stroke lesion segmentation using MRI data
[22]	2021	ML and DL for the prognosis and diagnosis of patients with large vessel occlusion stroke
[18]	2021	DL for stroke prognosis and diagnosis and overview of AI-based stroke imaging platforms
[19]	2020	ML and DL applied to stroke imaging and prediction of stroke-related outcomes
[23]	2020	ML for stroke prevention, diagnosis, treatment, and prognosis
[20]	2020	DL for stroke detection and lesion segmentation using neuro-imaging data
[21]	2019	ML for ischemic stroke diagnosis using neuro-imaging data and overview of AI-based software for large vessel occlusion stroke

TABLE II
OVERVIEW OF EXISTING LITERATURE REVIEWS RELATED TO MULTIMODAL LEARNING IN HEALTHCARE USING MEDICAL IMAGES, ELECTRONIC HEALTH RECORDS (EHR), AND OTHER MODALITIES

Ref.	Year	Scope
[27]	2022	Multimodal fusion using EHR data and medical images, with no restriction on type of disease
[30]	2022	Fusion of multimodal medical images, with no restriction on the types of imaging modalities or disease
[28]	2022	Multimodal learning using clinical, imaging, and omics data, with no restriction on type of disease
[32]	2021	Multimodal fusion of EHR data, including both structured and free-text data
[33]	2021	Multimodal fusion of medical signals from sensors within smart healthcare systems
[31]	2020	Multimodal fusion of neuro-imaging data for neurological disease
[29]	2020	Multimodal fusion of EHR and medical imaging data using deep learning

uses a numeric scale ranging from 0 to 6, where 0 indicates no symptoms and 6 refers to patient mortality [12]. Despite its wide adoption, mRS suffers from inter-observer reliability [13]. Stroke prognosis is affected by other factors, such as age, severity, time since onset, stroke location, and intracranial adverse events [14]. The variability of these factors across different patient groups increases the complexity of prognosis.

Table I provides an overview of the most relevant literature reviews on the use of machine learning for stroke-related tasks. In recent years, there has been a significant increase in the use of machine learning and deep learning methods for stroke detection and prognosis. This can be attributed to the advent of large-scale, digitized clinical data of various types, along with remarkable advancements in artificial intelligence algorithms and applications. Associated with this rise, a significant number of review articles were published between 2019 and 2022. In terms of the types of data modalities considered in these reviews, [15], [16], [17], [18], [19], [20], [21] focused on studies involving medical imaging only, while [22], [23] did not restrict their review to any type of clinical data. In general, most papers focus on medical imaging modalities, while clinical and omics data are peripherally considered.

Stroke prognosis and diagnosis are indeed multimodal in nature. In particular, clinicians rely on several sources of information in their decision-making process, including patient history, different types of medical imaging, omics data, and laboratory test results. Multimodal machine learning refers to the modeling of different but related data modalities using machine learning and deep learning [24]. Compared to learning with a single modality, multimodal learning approaches have proven to yield better results for various medical tasks as shown in recent work [25], [26].

Table II summarizes the scope of literature reviews related to multimodal learning in healthcare. Existing literature reviews focus on the use of Electronic Health Records (EHR), medical imaging and omics data [27], [28], [29], multimodal medical imaging, [30], [31], structured and free-text EHR data [32], and smart healthcare systems data [33] for a variety of clinical prediction tasks. Overall, existing literature reviews related to multimodal learning in healthcare have a broad scope and do not specifically target stroke prognosis and diagnosis, which is the primary focus of this paper.

The goal of this systematic review is to understand the scope of multimodal learning methods used in the context of stroke prognosis and diagnosis. We also investigate the types of multimodal datasets utilized and understand the expected clinical impact of the multimodal frameworks. The remainder of this paper is organized as follows: Section II summarizes the literature search strategy, Section III discusses the different multimodal methodologies and datasets, and finally Section IV describes opportunities for future work to advance multimodal learning for stroke prognosis and diagnosis.

II. PRISMA FRAMEWORK

A. Literature Search and Selection Strategy

We registered a protocol for this systematic review at the International Prospective Register of Systematic Reviews (PROSPERO) (registration ID: CRD42023390730) [34]. All of the relevant dates for article selection are documented in the registered protocol.

The literature search and selection process was conducted in accordance with the PRISMA 2020 guidelines [35]. For the inclusion and exclusion criteria, we included papers published between January 2015 and March 2023 in the English language since this corresponds to the period when multimodal machine learning in healthcare garnered more attention, as shown in Fig. 1. We included articles that were directly related to ischemic stroke, hemorrhagic stroke, or transient ischemic attack, addressed a prognostic or diagnostic task, and were published as original research papers. We only included articles that were published in journals categorized as Q1 by the Scopus Journal Ranking, at the time of performing the search. We excluded workshop papers and conference abstracts and ensured that the papers described the development or application of a multimodal framework that utilizes machine learning and/or deep learning.

To conduct the literature search, we identified a set of 51 relevant search terms based on our inclusion criteria. This set of keywords is organized into three categories, including *stroke*, *multimodal learning*, and *artificial intelligence*, with 17 keywords on average per category. Table III shows a subset of the selected search terms for each category. The complete set of terms per search category are provided in the registered

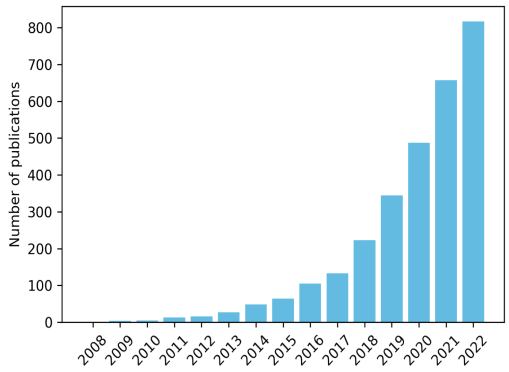

Fig. 1. Annual number of published articles on *PubMed* related to multimodal machine learning during the period of 2008 to 2022.

TABLE III
EXAMPLES OF SEARCH TERMS PER CATEGORY

Category	Search terms
Stroke	stroke, ischemic stroke, hemorrhagic stroke, stroke diagnosis, thrombus
Multimodal	multimodal alignment, multimodal fusion, data integration, joint fusion
Artificial intelligence	multi-layer perceptron, convolutional neural network, autoencoder, decision trees

protocol. We considered primary and secondary sources of information. The final list of primary sources included four scholarly databases, i.e., *PubMed, IEEE Xplore, ACM Digital Library*, and *Web of Science Core Collection*. For *PubMed* and *IEEE Xplore*, the search strategy included combinations of index terms and keywords, whereas the search strategy for *ACM Digital Library* and *Web of Science Core Collection* was based on combinations of keywords. For each primary source, we focused our search on the title, abstract, and keywords of potential articles. We then conducted a second round of manual searches on the references lists of all included papers to further identify the relevant papers.

To assess for eligibility, all retrieved papers underwent the same screening procedure. In particular, we removed duplicate articles and examined titles and abstracts. We then performed a full-text review to further refine the results. We used the *Covidence* software as our systematic review management tool to conduct the eligibility criteria assessment.

B. Data Collection and Extraction

For each included article, we extracted 24 data items: (1) first author, (2) publication year, (3) publication venue, (4) stroke type, (5) research objectives/contribution, (6) name of the proposed model, (7) types of data modalities, (8) details of data modalities, (9) source of the dataset, (10) dataset availability, (11) dataset name, (12) dataset split/ratios, (13) validation strategy, (14) value of K for K-fold cross-validation (if applicable), (15) test strategy, (16) multimodal learning paradigms used, (17) type of fusion (if applicable), (18) type of machine learning used (i.e., statistical machine learning or deep learning), (19) base models used, (20) nature of the stroke management task (prognosis or diagnosis), (21) exact name of the stroke management task, (22) nature of the downstream task, (23) reported performance metrics and (24) best-reported performance metrics. Note that the performance metrics depend mainly on the downstream task

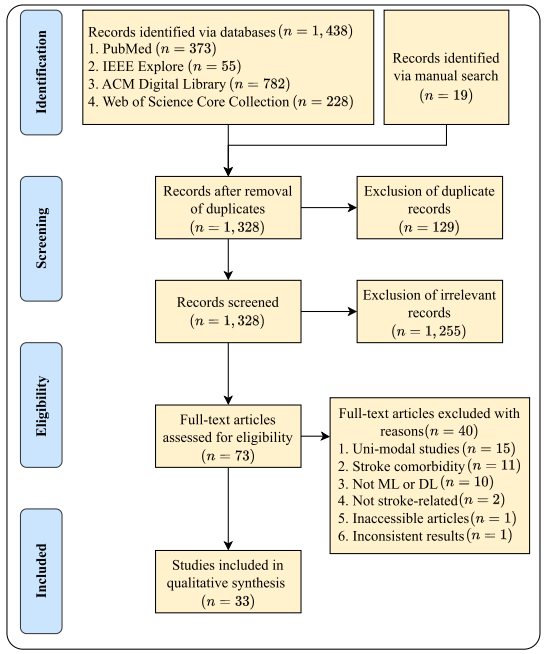

Fig. 2. PRISMA flowchart summarizing literature search and selection.

(e.g., classification or segmentation). We extracted the performance metrics for the best-performing models per article and ignored the results of any baseline models considered by the work.

C. Quality Assessment

We adopted two established tools for risk of bias and applicability assessment. We used Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2) [36] for diagnostic studies and Quality Assessment of Prognostic Accuracy Studies (QUAPAS) [37] for prognostic studies. We tailored both tools for the purpose of our study, by adjusting the signaling questions to fit the scope of our review.

III. RESULTS

A. Study Selection and Characteristics

We summarize the results of the literature search and selection strategy in Fig. 2, where n represents the number of articles at each stage. In the literature identification phase, we retrieved a total of 1,457 records. We excluded 129 duplicated articles, resulting in 1,328 articles for the screening phase. In the title and abstract screening phase, we excluded 1,255 records due to lack of relevancy. We further excluded 40 articles in the full-text eligibility assessment step. As a result, the systematic search and selection process resulted in 33 articles satisfying the eligibility criteria for further analysis.

Of the 33 included articles, 24 were journal publications, while ten were published as part of conference proceedings. Most papers focused on ischemic stroke ($n = 28$), with only one paper dealing exclusively with hemorrhagic stroke. Four papers addressed both ischemic and hemorrhagic strokes. Concerning the stroke management task addressed, most studies ($n = 30$) focused on stroke prognosis while the remaining ($n = 3$) developed multimodal models for stroke diagnosis.

TABLE IV
COMMONLY-USED ABBREVIATIONS IN THIS PAPER

Abbreviation	Full description
ADC	Apparent Diffusion Coefficient
AIF	Arterial Input Function
ASPECTS	Alberta Stroke Program Early Computed Tomography Score
CBF	Cerebral Blood Flow
CBV	Cerebral Blood Volume
CPSS	Cincinnati Pre-hospital Stroke Scale
CT	Computed Tomography
CTA	Computed Tomography Angiography
CTP	Computed Tomography Perfusion
DSC	Dynamic Susceptibility Contrast
DWI	Diffusion-Weighted Imaging
E-CTA	Electronic Computed Tomography Angiography
E-ASPECTS	Electronic Alberta Stroke Program Early Computed Tomography Score
ECASS	European Cooperative Acute Stroke Scale
ECG	Electrocardiogram
EEG	Electroencephalogram
FAST	Face Arm Speech Test
FLAIR	FLuid Attenuated Inversion Recovery
HMCAS	Hyperdense Middle Cerebral Artery Sign
EHR	Electronic Health Records
MRI	Magnetic Resonance Imaging
mRS	modified Rankin Scale
mTICI	Modified Thrombolysis in Cerebral Infarction
MTT	Mean Transient Time
NCCT	Non-Contrast CT
NASCET	North American Symptomatic Carotid Endarterectomy Trial
NIHSS	National Institutes of Health Stroke Scale
PPG	Photoplethysmography
PWI	Perfusion-Weighted Imaging
QOL	Quality of Life
RCBF	Relative Cerebral Blood flow
RCBV	Relative Cerebral Blood Volume
T1-c	T1-contrast
T1-w	T1-weighted
T2-w	T2-weighted
TFE	Turbo Field Echo
TICI	Thrombolysis In Cerebral Infarction
TIMI	Thrombolysis In Myocardial Infarct
Tmax	Time to maximum
TSE	Turbo Spin Echo
TSS	Time Since Stroke
TTI	Time To Imaging
TTP	Time To Peak
TTT	Time To Treatment

B. Datasets and Data Modalities

First, we summarize the main characteristics of the multimodal datasets, which are important for contextualizing the multimodal methodologies. We note that 12 studies used private datasets, 12 studies used publicly available datasets from community challenges, seven studies used clinical trials data, and three articles used public and private datasets. In terms of dataset size, they ranged between 28 and 574 patients. A full list of abbreviations of stroke-related modalities is provided in Table IV for ease of reference.

We summarize the main characteristics of the publicly available datasets in Table V. The most commonly used public datasets were released by the International Conference on Medical Imaging Computing and Computer-Assisted Intervention (MICCAI) as part of the annual Ischemic Stroke LEsion Segmentation (ISLES) challenge between 2015 and 2018 ($n = 12$). ISLES 2015 [38] consists of two datasets, namely, Subacute Ischemic Stroke lesion Segmentation (SISS) and Stroke outcome/Penumbra EStimation (SPES). While the former is multi-center, the latter is a single-center dataset. ISLES 2016-2018 [39], [40], [41] are all multi-center datasets. The main purpose of all the challenges is to provide benchmark datasets for the stroke lesion segmentation task. In these datasets, multimodal MRI is the most frequent data type, including diffusion and perfusion modalities. ISLES 2016 [39] is the only version that provides clinical variables to improve stroke lesion delineation, and to address the second goal of the challenge, which is prediction of functional outcomes in stroke patients. Overall, those datasets mainly provide multimodal medical imaging (CT or MRI), are tailored towards a very specific objective (mostly stroke lesion segmentation), are very small in terms of size (20–40 patients), and target a single stroke type (ischemic stroke).

The clinical trial datasets typically provide imaging and clinical data ($n = 7$). Imaging data include either CT [42], [43], [44], MRI [45], [46], or both [47], while clinical data are mostly clinical scores of measurable outcomes, such as NIHSS and mRS. MR CLEAN [42] is the only trial dataset that provides additional clinical data such as patient's medical history and demographics. As shown in Table V, two clinical trials collected their data from a single clinical institution [45], [46], while five combined data from multiple hospitals [42], [43], [44], [47]. Despite the clinical trial datasets being more comprehensive as they cover more diverse data modalities and larger sample sizes, access to them is generally restricted. The private datasets collected and used by some studies, which are usually of small size, are also not readily accessible.

We identified four main categories of data modalities used, including imaging, clinical, time-series, and other types of data, as summarized in Fig. 3. Imaging data includes both raw images and hand-crafted features extracted from images, and were considered by most of the included studies ($n = 30$). MRI is the most prevalent imaging modality ($n = 16$), which includes MRI Perfusion-Weighted Imaging (PWI) maps such as Cerebral Blood Flow (CBF), Cerebral Blood Volume (CBV), Relative CBF (RCBF), Relative CBV (RCBV) Mean Transient Time (MTT), Time to Peak (TTP), Time to Maximum (Tmax), Diffusion-Weighted Imaging (DWI), Apparent Diffusion Coefficient (ADC), T1-weighted (T1-w), T2-weighted (T2-w), and FLuid Attenuated Inversion Recovery (FLAIR). The second most frequent imaging modality is CT ($n = 14$), including Non-Contrast CT (NCCT), CT Perfusion (CTP), and CT Angiography (CTA). Some studies combined different imaging modalities, using them either as independent modalities [48], [49], [50], [51] or in conjunction with other non-imaging modalities, such as clinical records [52], [53], [54], [55].

Commonly used clinical data ($n = 14$) include patient demographics, vital signs, lab test results, past medical history, clinical measures, and quantitative image biomarkers. Patient age and sex are the most frequently used demographic variables ($n = 8$). Vital signs ($n = 5$) include measurements of blood pressure, blood oxygen, pulse rate, and temperature. Lab test results ($n = 3$) provide measurements of glucose and cholesterol levels, blood counts, and urinalysis. Medical history ($n = 6$) mainly provides information about past illnesses and the presence of chronic diseases, such as hypertension, diabetes mellitus, or cardiac disease. Other clinical measures ($n = 11$) include time-related variables such as Time since Stroke (TSS), Time to Treatment (TTT), and Time to Imaging (TTI), as well as stroke severity scoring scales such as mRS, NIHSS, and Thrombolysis in Cerebral Infarction (TICI). Lastly, some work used specific imaging biomarkers such as the Alberta Stroke Program Early CT Score (ASPECTS), infarct volume, penumbra volume, and side of the stroke.

Two of the included studies incorporated time-series data as input modalities. One study combined Electrocardiogram (ECG) and Photoplethysmography (PPG) data for stroke diagnosis [56], while another study used Electroencephalogram (EEG) for stroke prognosis [57].

A few studies used other data modalities ($n = 2$) that do not directly fit into any of the three main categories (imaging,

TABLE V
SUMMARY OF THE PUBLICLY AVAILABLE MULTIMODAL DATASETS RELATED TO STROKE, INCLUDING CLINICAL TRIALS DATASETS THAT MAY BE AVAILABLE UPON REQUEST

Name	Source	Type	Data modalities	Training set	Test set
ISLES 2015 SISS [38]	Single-center	Public challenge	Imaging: MRI (FLAIR, T2-w TSE, T1-w, TFE, TSE, DWI)	28	36
ISLES 2015 SPES [38]	Multi-center	Public challenge	Imaging: MRI (T1c, T2, DWI, CBF, CBV, TTP, Tmax) Imaging: MRI (ADC, MTT, Tmax, TTP, RCBV, RCBF)	30	20
ISLES 2016 [39]	Multi-center	Public challenge	Clinical: TICI, mRS, TSS, TTT	35	40
ISLES 2017 [40]	Multi-center	Public challenge	Imaging: MRI (DWI, ADC, CBV, CBF, MTT, TTP, Tmax)	43	32
ISLES 2018 [41]	Multi-center	Public challenge	Imaging: MRI (CBF, MTT, CBV, Tmax) and CTP Imaging: NCCT and CTA	63	40
MR CLEAN trial [42]	Multi-center	Clinical trial	Clinical: demographics, medical history, NIHSS, ASPECT and mRS Imaging: MRI (DWI, FLAIR, T2, DSC-PWI, follow-up FLAIR)	500	-
HIBISCUS trial [45]	Single-center	Clinical trial	Clinical: mRS, ECASS, NIHSS, QOL	233	-
I-KNOW trial	Multi-center	Clinical trial	Imaging: MRI (DWI, PWI) Imaging: multi-phase CTA	62	-
PRoVe-IT trial [43]	Multi-center	Clinical trial	Clinical: NIHSS, 90-day NIHSS, % NIHSS, mRS, ECASS 2 Imaging: MRI (T2, DWI, TOF-MRA, FLAIR, PWI)	500	-
1000plus trial [46]	Single-center	Clinical trial	Clinical: NIHSS and mRS Imaging: NCCT and CTA	1200	-
ESCAPE trial [44]	Multi-center	Clinical trial	Clinical: mRS Imaging: CT or MRI	500	-
HERMES trial [47]	Multi-center	Clinical trial	Clinical: mRS, NIHSS	1287	-

Full list of abbreviations for the data modalities is provided in Table IV.

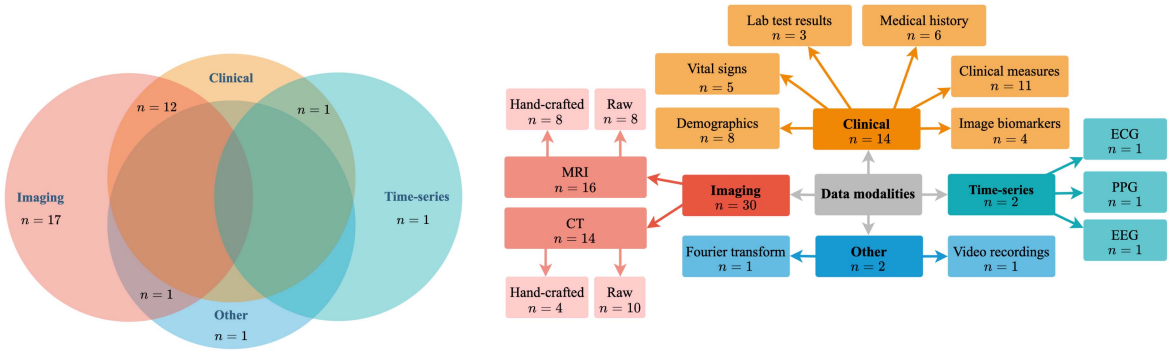

Fig. 3. Overview of data modalities used for stroke prognosis and diagnosis. (left): Venn diagram showing the intersection between data modalities as well as the number of papers that used each combination, denoted by n. (right): breakdown of the individual data modalities.

clinical, and time-series data). For example, one study collected facial video recordings for stroke patients in the emergency room. They derived video frames and their respective audio spectrograms to diagnose stroke patients [8]. One study used the phase and magnitude of the Fourier transform of CT scans as additional input modalities along with the original CT scan [58].

Overall, we observed that existing studies mostly use medical imaging and clinical data as inputs for their applications pertinent to stroke prognosis and diagnosis. Other types of data modalities are less explored, which motivates investigating their use and significance in relation to stroke-related tasks.

C. Multimodal Learning Methodologies

We present methodological advances of multimodal learning for stroke prognosis and diagnosis based on five commonly known core challenges in multimodal learning, namely, representation, alignment, translation, fusion, and co-learning [24], illustrated in Fig. 4. We note that M indicates a data modality, f is a modality-specific encoder, $+$ represents concatenation, $\sim$ represents a similarity operator, and $\sum$ represents an aggregation operator.

Early fusion approaches concatenate the multimodal features at the input level, joint fusion approaches perform feature concatenation at the representation level, and late fusion approaches aggregate modality-specific predictions. In joint representation learning, feature representations from both modalities are projected to the same space, while in coordinated representation learning, the representations of all modalities are projected into coordinated spaces that maintain a similarity measure across all modalities. Generative translation approaches use a generative model to map one modality to another. Example-based translation uses a dictionary to map one modality to another. Explicit alignment approaches align the corresponding sub-components among the instances of all modalities using a similarity measure. Implicit alignment approaches establish alignment in the latent space, such as via an attention module. In parallel co-learning, there is complete correspondence across the instances of all modalities. In non-parallel co-learning, correspondence is available between data categories instead of individual instances. In hybrid co-learning, a bridge modality is used to create a pairing between the main modalities.

Most studies included within the scope of this review focused only on multimodal fusion ($n = 18$), followed by the combination of multimodal representation learning and fusion ($n = 15$). Only one study addressed multimodal translation and fusion, while alignment and co-learning were not considered in any of the included papers. We also note that most studies used deep neural networks ($n = 20$), compared to using statistical machine learning models ($n = 13$), and only one article compared both modeling approaches. We also present detailed study characteristics in Table VI for prognosis-related tasks, and Table VII for diagnosis-related tasks. In the following sections, we discuss the

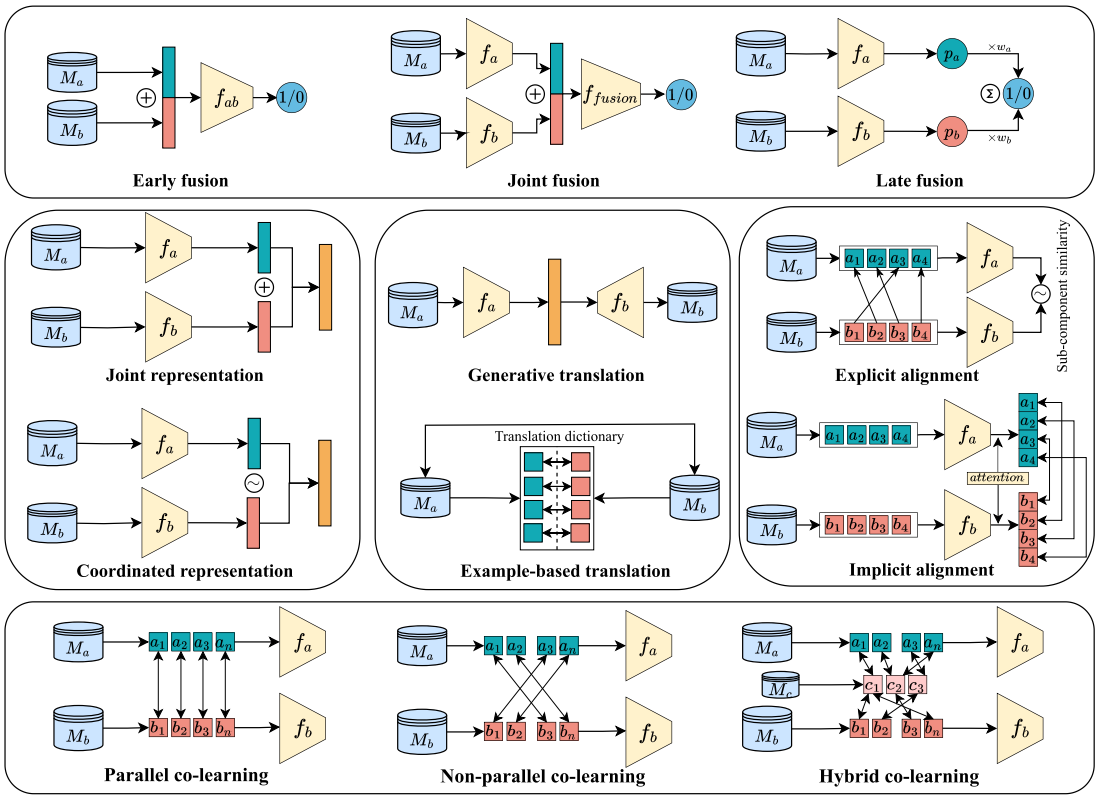

Fig. 4. Overview of the main multimodal learning paradigms.

methodologies proposed for stroke prognosis and diagnosis using each multimodal learning paradigm, including the reported performance results on the test set, where applicable.

1) Multimodal Fusion: Multimodal fusion refers to the process of combining information from multiple modalities for a downstream prediction or classification task, as opposed to relying on a single data modality, i.e. a single source of information. Based on the stage in which the features are combined, existing approaches can be categorized into early, joint, or late fusion [24], [29]. Any combination of these fusion methods can be considered a hybrid fusion approach. Early fusion is the most common type of fusion used among the studies included within the scope of this review ($n = 19$), followed by joint fusion ($n = 13$), and late fusion ($n = 3$). We highlight that existing fusion strategies rely on either statistical machine learning, deep neural networks, or both.

Early fusion with statistical machine learning: Early fusion approaches combine information from different modalities in the form of low-level features, such as via concatenation or pooling, which are then fed as a single input into a prediction model. Several studies used statistical machine learning models ($n = 12$), such as Random Forest (RF), eXtreme Gradient Boosting (XGboost), Light Gradient Boosting Machine (LightGBM), Categorical Boosting (CatBoost), and Support Vector Machine (SVM) or Support Vector Regression (SVR) to process the combined features.

In several studies, the authors derived features from different imaging modalities and concatenated them as a single input vector [59], [62], [66], [69], [70]. Some studies mainly considered radiomic features extracted from multimodal brain imaging [59], [66]. In another study, the authors focused on radiomic features extracted at the lesion area from five MRI modalities [59]. They compared the performance of five machine learning models (SVM, RF, LightGBM, Catboost, and XGboost) to predict mRS. Other work focused on radiomic features extracted from CT and NCCT to predict the probability of successful recanalization with intravenous alteplase using linear SVM model [13]. The authors further aimed at determining the best set of radiomic features characterizing thrombus in comparison to clinical thrombus characteristics. Both works evaluated their models via five-fold cross validation, and did not report model performance on an independent test set or provide any unimodal results for comparison.

One study introduced a two-stage model for stroke lesion segmentation [62]. The model concatenates intensity-based statistical features, template-based asymmetric features, and tissue probability feature maps extracted from four MRI modalities, namely, T1, T1-contrast (T1-c), T2, and FLAIR. Firstly, the authors trained a RF classifier using these features to identify the region of interest of the lesion from the whole brain image. Secondly, they trained a voxel-wise RF classifier to delineate the lesion. They reported improved sensitivity (0.750), precision (0.430) and dice coefficient (0.480) in the ISLES 2015 challenge [38] compared to top-ranked teams. No unimodal results were reported.

In other work, authors developed a model known as FASTER [69] that is built upon their previous model known as Segmentation Forest [80]. FASTER is an RF-based algorithm for automated volume estimation of reversibly damaged brain tissue, also known as penumbra. It comprises two classification

TABLE VI
SUMMARY OF THE CHARACTERISTICS OF THE INCLUDED PAPERS FOCUSING ON STROKE PROGNOSIS

Ref.	Year	Stroke	Task	Data modalities	Modeling	Validation	Testing	Results
Early fusion								
[59]	2023	Ischemic	Functional outcome prediction (mRS)	Hand-crafted: radiomic features extracted from multimodal MRI (DWI, ADC, FLAIR, SWI and T1-w)	Non-linear SVM, Random Forest, LightGBM, CatBoost, and XGBoost	Stratified 5-Fold CV	N/A	N/A
[60]	2021	Ischemic	Neurological deficit outcome prediction (NIHSS)	Clinical: demographics (age, gender) and medical history (modifiable and non-modifiable risk factors, blood pressure, glucose, hematocrit, hypertension, diabetes, smoking status, hyperlipidemia, atrial fibrillation, pre-treatment NIHSS score) Hand-crafted: features extracted from MRI or CT (FLAIR-NCCT atlas)	SVR with RBF kernel	N/A	Internal	MAE: 3.55 RMSE: 4.34 R^2: 0.43
[61]	2021	Ischemic	Stroke tissue outcome prediction	Clinical: time-related variables (stroke symptom onset-to-imaging and imaging-to-reperfusion time) Hand-crafted: features extracted from CTP average maps, Tmax, CBF, and CBV	RF with 80 trees and tree depth of 10	5-Fold CV	External	Accuracy: 0.898 Recall: 0.727 Specificity: 0.930 AUROC: 0.810 Dice: 0.388 VD: -3.1
[62]	2020	Ischemic	Stroke lesion segmentation	Hand-crafted: intensity-based statistical features (mean, minimum, maximum, variance, standard deviation, skewness, kurtosis, entropy) extracted from MRI (T1, T1-c, T2, FLAIR), template based asymmetry features (template difference, contralateral difference) extracted from MRI (T1, T1c, T2, FLAIR) and Gaussian mixture model based probability features extracted from MRI (T1, T1c, T2, FLAIR)	Cascaded RF with 200 trees and dense conditional random field, Hyperparameters are optimized using random search	5-Fold CV	Internal	Precision: 0.430 ± 0.290 Recall: 0.750 ± 0.320 Dice: 0.480 ± 0.290 HD: 45.9 ± 29.8 ASSD: 11.7 ± 16.0
[52]	2020	Ischemic	Functional outcome prediction (mRS)	Clinical: Demographics (age, sex), medical factors (pre-stroke mRS, initial NIHSS score, presence of wake-up stroke, duration from symptom onset to initial imaging, duration to groin puncture, intravenous r-tPA admission), existing health conditions (hypertension, hyperlipidemia, coronary heart disease, diabetes), angiographic intervention specifics (duration from groin puncture to recanalization, level of recanalization), post-intervention clinical indicators (NIHSS score post 24 hours), and additional clinical assessments (HbA1c, glucose) Hand-crafted: measures derived from CT scans (E-ASPECTS, acute ischemic volume), location of occluded vessel on CTA, collateral score from E-CTA, and measures from CTP (infarct core volume, ischemic penumbra volume, core/penumbra volume ratio)	GBM	Bootstrapping (63.2% of the training set)	Internal	Accuracy: 0.804 AUROC: 0.856
[57]	2020	Ischemic	Stroke effective recovery prediction	Clinical: NIHSS at acute phase Hand-crafted: features extracted from EEG signals (effective powers computed in 4 frequency bands and 22 bipolar channels)	Hand-crafted: Partial Least Squares Fusion: SVR with RBF kernel	Nested Leave-One-Out CV	N/A	N/A
[63]	2020	Ischemic	Stroke lesion segmentation	Imaging: MRI (T1, T2, FLAIR, DWI)	Ensemble of U-Net with multiple deconvolutional pathway	5-Fold CV	N/A	N/A
[64]	2019	Ischemic	Stroke lesion segmentation	Imaging: MRI (T2, DWI)	Res-U-Net	Leave-P-Out CV ($P = 25$)	External	Dice score: 0.890 HD: 1.51
[65]	2019	Ischemic	Functional outcome prediction (mRS)	Clinical: Demographics (age, sex) and admission details (total/specific ASPECTS, collateral status rating, presence of HMCAS, volume of infarction on CTP, volume of penumbra on CTP, stroke site, occluded vessel site, initial TIMI score, degree of stenosis according to NASCET on left and right hemispheres, baseline NIHSS score, NIHSS score after 24 hours)	XGBoost GBM	Stratified 5-Fold CV	N/A	N/A
[66]	2019	Ischemic	Recanalization prediction with intravenous alteplase	Hand-crafted: radiomics features derived from NCCT and CTA (first-order statistics, shape and size, textural features)	Linear SVM	5-Fold CV	N/A	N/A
[67]	2018	Ischemic	Functional outcome prediction (mRS)	Clinical: TSS, TTT, TICI Hand-crafted: features extracted from ADC, MTT, RCBF, RCBV, Tmax, and TTP MRI images	C-SVM with RBF kernel	Leave-One-Out CV	External	AUROC: 0.934
[68]	2018	Ischemic	Stroke tissue outcome prediction	Clinical: TSS, TTT, mRS score, and TICI score (population- and patient- level) Imaging: MRI ADC map and MRI perfusion maps (RCBF, RCBV, MTT, TTP, Tmax)	Clinical: Gated Recurrent Unit (GRU) Imaging: U-Net Fusion: FCNN	7-Fold CV	Internal	Precision: 0.260 ± 0.230 Recall: 0.610 ± 0.350 Dice: 0.290 ± 0.220 HD: 47.2 ± 22.1 ASSD: 7.20 ± 4.14
[69]	2017	Ischemic	Stroke lesion segmentation	Hand-crafted: local and global features extracted from MRI (T1-w, T2-w, DWI, DSC)	Segmentation forest	N/A	Internal	Precision: 0.560 Recall: 0.530 Specificity: 0.990 Dice score: 0.340
[48]	2021	Ischemic	Ischemic core/deficit segmentation	Imaging: CTP phases (NCCT, CTA, CTA+)	3D U-Net	4-Fold CV	External	Dice (core): 0.610 Dice (deficit): 0.680 R (core): 0.800 R (deficit): 0.67
[49]	2019	Ischemic	Ischemic core/deficit segmentation	Imaging: CT and CT perfusion maps (CBF, CBV, MTT, Tmax)	2D U-Net randomly initialized, Hyperparameters optimized using random search	5-Fold CV	Internal	Recall: 0.570 ± 0.350 Dice: 0.490 ± 0.310 HD: 11.3 ± 31.6 PPV: 0.510 ± 0.360
[70]	2018	Ischemic	Ischemic core/deficit segmentation	Hand-crafted: features extracted from MRI imaging using: (1) concentration curves: CBV and TTP, (2) oscillatory singular value decomposition deconvolution: CBV, MTT, and Tmax, (3) statistical (parametric) approach: parametric CBV, parametric MTT, parametric Tmax, relative transit time heterogeneity, capillary transit time heterogeneity, and cerebral metabolic rate of oxygen	XGBoost	Leave-One-Out CV 5-Fold CV	N/A	N/A
[50]	2018	Ischemic	Ischemic core/deficit segmentation	Imaging: Multi-spectral MRI (DWI, T2-w, ADC)	2D U-Net with ResNet50 backbone, randomly initialized.	Training split (10%) 2-Fold CV	Internal External	Precision: 0.610 Recall: 0.600 Dice: 0.550 HD: 51.0 ASSD: 9.21
Early fusion and translation								
[51]	2019	Ischemic	Ischemic core/deficit segmentation	Imaging: CT perfusion maps (CBF, CBV, MTT and Tmax) and DWI MRI	Imaging: CGAN Fusion: Fully Convolutional Neural Network (FCNN)	5-Fold CV	N/A	N/A

CONTINUED

Ref.	Year	Stroke	Task	Data modalities	Modeling	Validation	Testing	Results
Joint fusion and representation learning								
[71]	2022	Ischemic	Ischemic core/deficit segmentation	Imaging: 4D CTP maps (CBV, CBF, Tmax, MTT)	Imaging: U-Net Fusion: Attention block and residual paths with CNN layers, randomly initialized	Training split (N/S)	Internal	Precision: 0.480 Recall: 0 Dice: 0.480 HD: 18.5 F_1-score: 0.530 F_2-score: 0.560
[72]	2021	Ischemic	Stroke tissue outcome prediction	Hand-crafted: features extracted from DSC MRI Imaging: perfusion maps from DSC-MRI (Tmax, TTP, MTT, RCBF, RCBV, ADC)	Hand-crafted: 2D U-Net Imaging: 2D U-Net Fusion: FCNN	10-Fold CV	Internal	Precision: 0.290 ± 0.230 Recall: 0.630 ± 0.300 Dice: 0.310 ± 0.210 HD 33.917.4 ASSD 5.99 ± 4.58
[73]	2021	Hemorrhagic	Early hematoma expansion prediction	Clinical: demographics (age, gender) and medical history (diastolic blood pressure, systolic blood pressure, onset time and hematoma volume Imaging: 3D CT	Clinical: transposed CNN Imaging: CNN with channel attention Fusion: FCNN	4-Fold CV	N/A	N/A
[53]	2020	Ischemic	Functional outcome prediction (mRS)	Clinical: demographics (age, gender) and medical history (initial NIHSS, cardiac history, diabetes, hypercholesterolemia, thrombolysis treatment) Imaging: 3D acute Time of Flight-MRA	Clinical: FCNN Imaging data: CNN Fusion: FCNN Hyperparameters are optimized using random search	Training split: 20%	Internal	AUROC: 0.760
[54]	2020	Ischemic	Functional outcome prediction (mRS)	Clinical: treatment information (endovascular treatment) and clinical metadata (demographics, medical history, radiological image biomarkers (ASPECT score, occlusion site, collateral score and symptom side), medical vitals signs (hypertension, glucose level, etc.) at time of admission to hospital) Imaging: 3D NCCT	Clinical: FCNN Imaging: CNN with attention, randomly initialized. Fusion: FCNN Hyperparameters are optimized using random search	N/A	Internal	Accuracy: 0.770 AUROC: 0.750 F_1-score: 0.620
[55]	2019	Ischemic	Functional outcome prediction (mRS)	Clinical: demographics (age, gender) and medical history (time from stroke onset to presentation, NIHSS at presentation, blood pressure at presentation, blood glucose level at presentation, temperature at presentation, Medical history (hypertension, diabetes, hypercholesterolemia or atrial fibrillation) Imaging: NCCT	Clinical: FCNN Imaging: CNN Fusion: FCNN	10-Fold CV	Internal	Accuracy: 0.740 Recall: 0.560 Specificity.: 0.930 AUROC: 0.750 PPV: 0.900 NPV: 0.670 F_1-score: 0.690
[74]	2019	Ischemic	Stroke tissue outcome prediction	Clinical: time between stroke onset and CTP imaging, time between CTP imaging and the end of the mechanical thrombectomy, mTICI scale after thrombectomy, persistence of the occlusion at 24 hours CT angiography, and arterial input function Imaging: 4D CTP	Clinical: CNN and upsampling Imaging: CNN and upsampling Fusion: CNN	5-Fold CV	N/A	AUPRC: 0.540 Dice: 0.480 Soft Dice: 0.400 MAE: 36.7
[75]	2019	Ischemic	Stroke tissue outcome prediction	Imaging: 4D PWI, 3D MRI ADC map and 3D MRI perfusion maps (RCBF, RCBV, MTT, Tmax, TTP)	Imaging: U-Net with squeeze and excitation Fusion: CNN	N/A	Internal	Precision: 0.260 ± 0.230 Recall: 0.610 ± 0.280 Dice: 0.290 ± 0.210 HD: 38.9 ± 20.0 ASSD: 7.00 ± 4.31
[76]	2019	Ischemic, hemorrhagic	Stroke lesion segmentation	Imaging: MRI (T1-weighted, FLAIR)	Imaging: U-Nets Fusion: 3D CNN	5-Fold CV	External	Dice: 0.54
[77]	2018	Ischemic	Stroke tissue outcome prediction	Imaging: 4D PWI, 3D MRI perfusion maps (RCBF, RCBV, MTT, Tmax, TTP) and 3D MRI ADC map	Imaging: U-Net Fusion: GRU and CNN layers	Training split: 16%	Internal	Precision: 0.230 ± 0.210 Recall: 0.660 ± 0.290 Dice: 0.290 ± 0.210 HD: 41.6 ± 22.0 ASSD: 7.69 ± 5.71
Late fusion and representation learning								
[78]	2022	Ischemic	Functional outcome prediction (mRS)	Clinical: Age, NIHSS Imaging: diffusion MRI (raw, ADC) and MRI perfusion maps (Tmax, CBV, CBF)	Imaging: CNN-LSTM Fusion: Weighted decision using the clinical data Hyperparameters are optimized using grid search	5-Fold CV	N/A	N/A
[79]	2019	Ischemic	Ischemic core/deficit segmentation	Hand-crafted: features extracted from CT Imaging: CT perfusion maps (CBF, CBV, Tmax, MTT)	Hand-crafted and imaging: U-Net Fusion: classification models (voting, simple weighted averaging, and logistic regression)	10-Fold CV	Internal	Recall: 0.730 Dice: 0.710 VS: 0.820

TABLE VII
SUMMARY OF THE CHARACTERISTICS OF THE INCLUDED PAPERS FOCUSING ON STROKE DIAGNOSIS

Ref.	Year	Stroke	Task	Data modalities	Modeling	Validation	Testing	Results
Early fusion								
[56]	2022	Ischemic, hemorrhagic	Stroke prediction and semantic interpretation	Hand-crafted: features extracted from ECG and PPG biosignals Time-series: ECG and PPG biosignals	Naive Bayes, Logistic regression, MLP, RF, BFTree, BayesNet, AdaBoost, non-linear SVM, C4.5, LSTM, BiLSTM, CNN-LSTM	5/10/20-Fold CV	Internal	Accuracy: 0.953 Precision: 0.953 Recall: 0.953
Joint fusion and representation learning								
[8]	2022	Ischemic, hemorrhagic, TIA	Stroke diagnosis in the emergency room	Other: Audio spectrograms and video frames extracted from patient facial recordings	Video: ResNet-34, ImageNet pretrained Audio: ResNet-18, ImageNet pretrained Fusion: DCGAN	5-Fold CV	Internal, Reader study	AUROC: 0.721 (0.677, 0.790)
Late fusion and representation learning								
[58]	2021	Ischemic, Hemorrhagic	Stroke prediction and classification	Imaging: CT Other: Fourier transform components (magnitude and phase angle) extracted from CT	Imaging: Restricted Boltzmann Machines Other: Restricted Boltzmann Machines Fusion: softmax layer	N/A	Internal	Accuracy: 0.997 ± 0.004

stages, each with two RF classifiers. Firstly, FASTER delineates the stroke lesion to identify the region of interest using features extracted from different sets of MRI maps modeled with the classifiers. Following that, the classifiers predict the exact stroke lesion voxels. The authors concatenated statistical features extracted from T1-w, T2-w, DWI, and Dynamic Susceptibility Contrast (DSC) perfusion MRI modalities. In total, they extracted 274 features per voxel. Their model achieved 0.560, 0.530, 0.990, and 0.340 in terms of precision, recall, specificity, and dice coefficient, respectively. They did not compare the results of the proposed model to any unimodal baselines.

In another study, 12 parameters were derived from DSC-MRI perfusion maps, DWI, and FLAIR for infarct prediction, where two parameters were derived from DWI and FLAIR scans and the rest were perfusion parameters [70]. The authors used three methods to derive the perfusion maps, including oscillatory

singular value decomposition, a parametric (statistical) approach, and direct derivation of cCBV and TTP from the concentration curve. They used all parameters to train an XGBoost voxel-wise classifier for ischemic core segmentation via cross-validation. They compared its performance to that of a generalized linear model. The authors did not report performance on an independent test set and as compared to any unimodal baselines.

Other work used features extracted from time-series data [56], [57]. Specifically, one model was proposed to concatenate features extracted from ECG and PPG bio-signals for real-time stroke diagnosis [56]. Nine classical machine learning models were evaluated, and RF yielded the best classification performance. No unimodal results were reported. In another study, authors used features extracted from EEG data with the NIHSS score to predict effective recovery in stroke patients [57]. First, they trained a partial least square model on 22 bipolar EEG channels to compute the effective power of each bipolar channel. Following that, they used the extracted effective power measurements as input features along with the NIHSS score, via concatenation, to an SVR model. The authors did not report performance on an independent test set or baseline unimodal performance.

Several research studies explored the fusion of clinical data extracted from EHR, clinical measurements, and features extracted from medical imaging modalities [52], [60], [61], [65], [67]. For example, clinical data, including demographics, patient risk factors, and pre-treatment NIHSS score with biomarkers derived from CT or MRI, were integrated to predict the 30-days NIHSS score [60]. Lesion symptom mapping [81] and Rrelief [82] were used as feature selection methods to extract the biomarkers from the lesion's region of interest. The feature vector was used to train a support vector regression model. The best performance was achieved using the combination of clinical data and features extracted via Relief with a Mean Absolute Error (MAE) of 3.55 compared to 4.33 using clinical data.

In other work, authors trained gradient boosting models using patient demographics, clinical measures such as mRS and NIHSS, and biomarkers extracted from CT as input features to predict the mRS score [52], [65]. No results were reported on an independent test set [65]. In another study, the authors performed step-wise feature inclusion evaluation of four sets of features including (i) patient demographics, clinical measures, medical history, and conventional CT biomarkers (ii) advanced imaging characteristics from CTP, (iii) interventional characteristics, and (iv) post-interventional characteristics study [52]. The best performance was achieved using the full set of features from all modalities, reaching an accuracy of 0.800 and an AUROC of 0.860, outperforming the baseline model by 0.100 for both accuracy and AUROC.

Other methods were developed to integrate time-related measurements with features extracted from imaging data [61], [67]. The former combined stroke-onset to imaging and imaging to perfusion times with average maps derived from CTP including Tmax, cBF, and cBV for each voxel to predict infarct volume in acute ischemic stroke patients. All derived features were concatenated and passed to a voxel-wise RF classifier to predict whether the voxel is normal or part of the infarction. The model achieved an AUROC score of 0.810, with improvements of 0.600 and 0.100 compared to conventional methods used for Tmax and cBF maps, respectively. Similarly, in another study, the authors incorporated TSS, TTT, and TICI, with features extracted from ADC MRI and perfusion MRI maps [67]. They used an SVM model to predict mRS. The model achieved an average AUROC score of 0.934. No unimodal results were reported.

Early fusion with deep neural networks: Compared to the use of statistical machine learning methods in early fusion, seven papers focused on the use of deep neural networks. The majority ($n = 5$) stack multimodal CT or MRI scans as image channels at the input level [48], [49], [50], [51], [64], [83]. For CT imaging, one study proposed the fusion of NCCT, CTA, and CTA+ for the segmentation of ischemic core/deficit [48]. The stacked volume is then passed to a 3D U-Net segmentation model for core/deficit volume estimation. The proposed approach achieved accuracy and AUROC scores of 0.900 and 0.810, respectively, which is similar to the performance obtained using CTA only. The proposed approach achieves performance gains of 0.070 and 0.080 for accuracy and AUROC, respectively, compared to the NCCT-only model. In another study, authors proposed symmetric modality augmentation to merge CT and CTP maps prior to modeling [49]. Symmetric modality augmentation exploits the brain hemispheres' symmetry to learn robust features. In this regard, each CT scan and four perfusion maps, which were cBF, cBV, MTT, Tmax, were flipped over the mid-sagittal axis. Both versions were then concatenated to form a single volume, which was passed into a U-Net model to generate the lesion mask. This model achieved a dice coefficient of 0.490 on the ISLES 2018 [41] test set. No unimodal performance results were reported.

For MRI imaging, one study proposed a single-encoder multi-decoder U-Net framework for stroke lesion segmentation [63]. The authors stacked four MRI modalities, including T1, T2, FLAIR, and DWI, and passed them to the encoder network as a volume. They also enhanced the U-Net architecture with several blocks, such as multi-residual attention and multi-task heads to improve stroke lesion segmentation. Their model achieved a dice coefficient of 0.775 on the ISLES 2015 [38] SISS test set. Similarly, another study proposed combining DWI, ADC and T2-w MRI for stroke lesion segmentation, known as Res-FCN [50]. They introduced additional blocks, such as dilated convolutional, global convolution, and boundary refinement blocks to improve the segmentation quality and reduce false negatives. This model achieved a dice score of 0.645 on the internal test set and 0.550 on the ISLES 2015 test set (external test set). The authors did not provide unimodal results. In one study, DWI and T2 MRI scans were merged to improve lesion appearance for stroke lesion segmentation [64]. This fused image is then passed to the segmentation network along with the raw DWI scans. The multimodal model achieved a dice coefficient of 0.884 on the ISLES 2015 [38] SPES training set, which is used as an external test set, and 0.740 on the internal test set. The model outperformed unimodal baselines on both test sets by 0.040.

Joint fusion with deep neural networks: In joint fusion, a modality-specific neural network is used for each input modality to learn useful feature representations. The learned representations from each modality are then passed to a trainable fusion model to compute the predictions. The main advantage of joint fusion is that it enables end-to-end multimodal training which, in turn, enhances the quality of the learned representations from all modality encoders. Joint fusion ranks as the second most commonly utilized form of multimodal fusion ($n = 13$). All studies that used joint fusion relied on the use of deep neural networks. More specifically, ten studies proposed models that jointly learn modality-specific representations from each modality, and then

concatenate them as a single representation prior to the fusion module. The three remaining studies performed intermediate joint fusion within layers of convolutional neural networks.

One study presented an end-to-end multimodal framework that learns representations from clinical data, including patient demographics and medical history, and Time-of-Flight MR angiography simultaneously [53]. The authors used a 3D CNN and a Multi-Layer Perception (MLP) as feature extraction encoders for the images and clinical data, respectively. The learned representations from each modality were then fused via vector concatenation and fed into a fusion module to predict the mRS score. The authors reported an AUROC of 0.760 for the multimodal model, which shows 0.010 and 0.080 performance gains compared to the unimodal clinical data and imaging models, respectively. The authors compared their model performance to an early fusion baseline which achieved an AUROC of 0.750.

Other methods followed a similar architectural design but used 3D NCCT and clinical data to improve mRS prediction [54], [55]. Similar classification accuracy and AUROC values were reported, which were around 0.750 for both metrics. While one model managed to achieve performance gains compared to each of the unimodal CT and clinical data models (0.170 and 0.100 for AUROC and accuracy, respectively) [55], the other model performed better than the clinical data model and on par with the CT imaging model [54].

In one study, the authors used transposed convolutions to learn a volumetric representation of demographics and medical history data, and fused it over the channel dimension with representations of CT volumes extracted using a convolutional encoder to predict early hematoma expansion [73]. This was followed by average pooling and a 1×1 convolutional filter to transform the fused representations before computing the prediction. The authors did not report performance on a test set but reported a performance gain of nearly 0.100 on the validation set with respect to the unimodal image model. Similarly, one approach jointly modeled CTP map volumes along with Arterial Input Function (AIF) measurements and treatment metadata [74]. The framework incorporated an upsampling network to model treatment metadata and AIF measurements to learn a volumetric representation which is later merged with the CTP encoding on the channel dimension. The model resulted with an AUROC of 0.540, and roughly maintained the same performance when the treatment metadata variables were omitted.

A multi-branch model was also developed for 4D PWI and standard diffusion/perfusion MRI maps [72], [77]. Rather than modeling all of the PWI sequences, the model selects slices associated with higher contrasting agent concentration. The two modalities are then processed using separate U-Net models and the output of each network is then fed into a bi-directional Gated Recurrent Unit (GRU) layer [84]. Lastly, the representations of all modalities are merged on the channel dimension using a set of convolutional layers. The model achieved a precision and recall of 0.230 and 0.660 in one study [77], and 0.290 and 0.630 in another study [72]. Both studies showed a slight improvement over the unimodal models.

In another study [75], the authors extended the model in previous work [77] by incorporating a squeeze-and-excitation block following the fusion module. They reported a 0.030 improvement in precision. In the same context, another model was proposed that includes the TICI score at the patient and population levels to enhance lesion outcome prediction [68]. At the patient level, the TICI score is encoded as an extra channel as input to the penultimate layer of the network. At the population level, the TICI score is incorporated into the loss function as a weighting parameter. The authors adopted an existing architecture [77], but with a single branch rather than multiple branches. Their model achieved 0.290 on the ISLES 2017 [40] test set. The model achieved a 0.050 improvement in the dice score with respect to the unimodal model.

Several studies proposed to jointly fuse the feature maps at the convolutional block level [71], [76], such as the DeepStroke model for emergency room stroke diagnosis (triage) using facial video recordings to support patient triage [8]. Audio data is represented as audio spectrogram images and modeled using ResNet18, while video data is processed as the difference between two consecutive frames and modeled using ResNet-34. The authors used lateral connections to concatenate the feature maps learned from the audio spectrograms and video frames at the convolutional block level of each modality encoder. The output of the video frames encoder is then passed through a fully connected layer to make the final predictions. DeepStroke was able to achieve a diagnostic AUROC score of 0.721, which entails 0.200 and 0.250 improvements compared to the unimodal video and audio models, respectively.

A four-path model was also proposed to encode the CTP-based maps cBV, cBF, Tmax, and MTT for ischemic core segmentation [71]. Each modality is encoded via a modality-specific encoder, while the feature maps generated from the pairs (cBV, cBF) and (Tmax, MTT) were merged at the level of the convolutional layer. The output of the final layers of every pair is passed to a cross-modal cross-attention module that captures the interactions between modality pairs. The features were then passed to a decoder that outputs the segmentation mask. The proposed methodology was able to improve recall compared to the baselines, achieving 0.590. Furthermore, the cross-attention module brought significant improvements compared to the unimodal models using single maps as input, i.e. cBV, cBF, Tmax, and MTT. The improvement ranged from 0.070 to 0.150 for precision, recall, dice coefficient, and F_1 score.

In one study, the authors proposed a nine-path segmentation model that considers the axial, sagittal, and coronal planes in the MRI volume [76]. Each path consists of two encoders and receives the T1-w and FLAIR pair in case they are both present, or T1-w with its flipped version in case FLAIR is missing. The learned feature maps resulting from every convolutional layer for both modalities are merged via convolutional layers rather than concatenation. The authors reported better performance compared to using a single modality.

Late fusion with deep neural networks: Late fusion approaches combine predictions computed by modality-specific models at the output level. Fusion of the predictions is typically performed via weighting, majority voting, or averaging. Late fusion is the least used form of fusion in the included papers ($n = 3$), despite its simplicity. For example, it was used to fuse pixel-level predictions of four separate 2D U-Net architectures that learn from single CTP maps, including cBV, cBF, Tmax, and MTT [79]. Each pixel is classified as to whether it belongs to a healthy or stroke lesion based on different prediction ensembling techniques, such as average weighting, majority voting, and logistic regression applied to the four probability maps produced by the U-Nets. The logistic regression classifier achieved the best performance with 0.730, 0.710, and 0.820 for recall, dice coefficient, and volume similarity, respectively. The proposed multimodal model nearly doubled

the performance obtained using each of the four unimodal models.

Another study proposed a CNN-LSTM model to predict the functional outcome of stroke patients [78]. The model separately processes sequences of diffusion MRI (ADC and DWI) and perfusion MRI (cBV, cBF, and Tmax). The authors used the CNN to extract features from the input maps, and the LSTM for temporal modeling of the sequences. The encodings of the five modalities are then fused and weighted, either by the patient's normalized age or NIHSS, via ensemble voting to produce the final prediction. The authors did not report test set performance, but their model resulted with a 0.050 performance gain compared to the unimodal models when using the NIHSS score for weighting the individual predictions.

Late fusion with statistical machine learning: In one study, the authors proposed combining CT scans with their Fourier transform to predict the type of stroke using Restricted Boltzmann Machine [58]. The Fourier transform is applied to each image to obtain its phase and magnitude. The authors trained separate RBM models using the original CT scans with the phase and magnitude features. They computed the final predictions via a softmax layer by aggregating the representations from each modality. The best classification accuracy was obtained using raw CT scans with phase data, achieving an accuracy of 0.997, illustrating a better performance compared to using CT scans only, with 0.006 accuracy gain.

2) Multimodal Representation Learning: Representation learning refers to the process of converting the original data into a computationally-friendly format, such as vectors [85]. Learning representations from multiple modalities simultaneously is referred to as multimodal representation learning, and is often performed in conjunction with a downstream fusion task. Multimodal representation learning can be broadly categorized into joint and coordinated representation learning [24]. Joint representation learning approaches project the unimodal representation of each modality into the same latent space, and then they are combined via concatenation as an example. Coordinated representation approaches impose a similarity criterion on the representations of the modalities to bring them closer within the latent space [24]. Both methods can be implemented in a single latent space or at intermediate levels within the model. Among all the included papers, 14 papers considered multimodal representation learning [8], [53], [54], [55], [58], [71], [72], [73], [74], [75], [76], [77], [78], [79]. Though these papers did not discuss multimodal representation learning explicitly, they incorporated it as part of their methodological implementation based on the definition of multimodal representation learning. All of studies implemented joint representation learning within intermediate model layers [8], [71], [76], or in the latent space and/or prediction level [53], [54], [55], [58], [72], [73], [74], [75], [77], [78], [79].

3) Multimodal Translation: Multimodal translation refers to the process of mapping one data modality into another one. There are two types of translation approaches namely, example-based and generative translation [24]. Example-based translation approaches rely on a predefined look-up dictionary, while generative translation approaches develop models that learn the translation between modalities during training. Only one study used multimodal translation, specifically generative multimodal translation [51]. The authors translated CTP maps into DWI MRI using a CGAN. The primary purpose of the mapping was to improve the delineation of the hyper-intense regions in patients with stroke. Both modalities were then fused together at the input level over the channel dimension, and passed to a U-Net model for stroke lesion segmentation. Their proposed approach achieved a dice score of 0.540, which is slightly better than the model trained using CTP scans only, which achieved a dice score of 0.530.

D. Risk of Bias and Applicability Assessment

We summarize the results of the quality assessment using QUADAS-2 for diagnostic studies and QUAPAS for prognostic studies included in our systematic review in Fig. 5.

For diagnostic studies, one study showed a high risk of bias attributed to the lack of a proper description of the patient inclusion/exclusion criteria. In addition, all studies were identified as having a high risk of bias due to confining model testing to either internal test sets, or cross-validation without reporting results on external test sets, which may affect the generalizability of these models. Furthermore, one study showed a high risk of bias due to the lack of unimodal baselines. All studies were categorized as showing low risk of bias with respect to the flow and timing domain. Finally, concerns regarding applicability were low for all studies with respect to the scope of our review.

For prognostic studies, the vast majority of the included studies have a low risk of bias with respect to patient selection. However, three studies showed a high risk of bias due to the lack of an appropriate description of the inclusion/exclusion criteria, while two studies had an unclear risk of bias as exclusion criteria was not discussed. With respect to the index test, we found that 29 studies have a high risk of bias due to the lack of both internal and external testing. Only one study reported their results using internal and external test sets. Regarding the outcome, only one study showed an unclear risk of bias, while the remaining studies showed a low risk of bias. For flow and timing, 13 studies demonstrated a high risk of bias due to insufficient data leakage safeguards and temporal inconsistencies between modalities. For the analysis domain, three studies showed a high risk of bias since they limit their evaluation to external test sets. Lastly, we found all studies to show low concerns regarding applicability.

E. Clinical Applications

The papers considered in this study focus on either stroke prognosis (Table VI) or diagnosis (Table VII). Overall, 30 studies introduced stroke prognostic approaches focusing on functionality outcomes and brain tissue status. Eight studies used the 90-day mRS score to predict functional outcomes in stroke patients [52], [53], [54], [55], [59], [65], [67], [78]. Two papers focused on forecasting the neurological deficit outcome, where [60] predicted the 30-days NIHSS and [57] predicted effective recovery. They computed effective recovery based on the relative change of the NIHSS between stroke onset and 180 days after stroke. Concerning brain tissue status, six studies developed models to predict tissue outcome, specifically penumbra delineation, [61], [68], [72], [74], [75], [77]. Six studies performed stroke lesion segmentation based on follow-up imaging [62], [64], [73], [76], [80], [83]. Four studies proposed segmentation models using CTP maps [48], [49], [51], [71], [79] and two studies proposed segmentation models using MRI data [50], [70], to identify irreversibly damaged tissue prior to

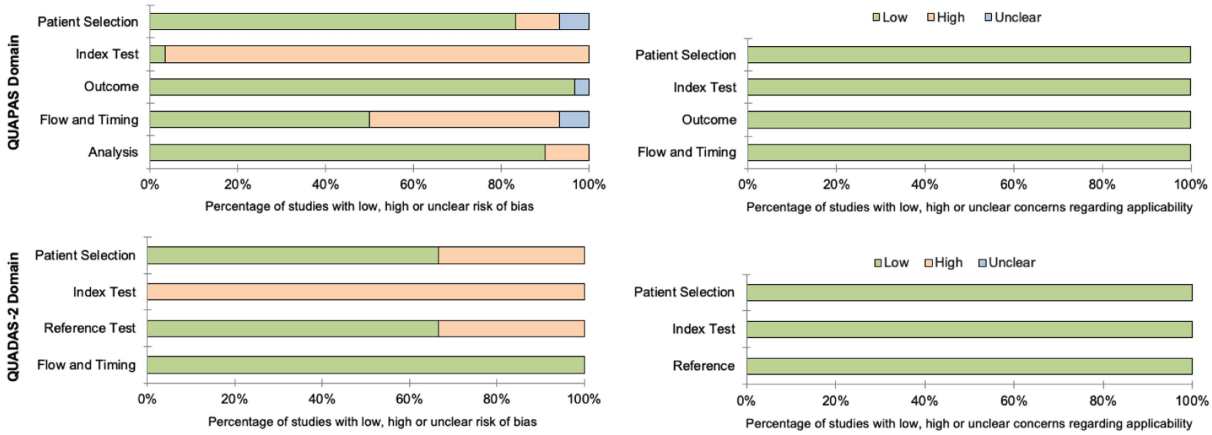

Fig. 5. Risk of bias and applicability results for QUAPAS and QUADAS-2.

TABLE VIII
SUMMARY OF ADVANTAGES AND DISADVANTAGES OF IMAGING MODALITIES

Modalities								
MRI modalities	**DWI**	**ADC**	**PWI**	**MRA**	**FLAIR**	**T2-Weighted**	**SWI**	**TTP**
Advantages	- Detects ischemic stroke within minutes of onset. - Differentiates between acute and chronic lesions. - Helps in estimating the infarct core size.	- Quantifies the level of water diffusion restriction, aiding in identifying the infarcted tissue. - Complements DWI by confirming the presence of ischemic stroke and distinguishing it from other types of brain lesions. - Helps in monitoring the evolution of ischemic stroke over time.	- Identifies tissue at risk of infarction but still salvageable (penumbral). - Differentiates between benign oligemia and true penumbra. - Guides therapeutic decisions, such as the use of thrombolytic agents.	- Non-invasive assessment of intracranial and extracranial vessels. - Identifies vessel occlusion or stenosis contributing to the stroke. - Helps in planning interventional procedures like thrombectomy.	- Identifies acute and subacute stroke lesions by visualizing hyperintensities. - Differentiates between new and chronic white matter changes. - Helps in assessing lesion age by comparison with DWI.	- Identifies areas of hemorrhage and microbleeds. - Assists in differentiating between hemorrhagic and ischemic stroke. - Evaluates the risk of hemorrhagic transformation in ischemic stroke.	- Highly sensitive to blood products, including microbleeds and hemorrhages. - Can detect small venous structures and blood oxygenation levels. - Useful for identifying hemorrhagic transformation, cerebral microbleeds, and vascular malformations.	- Helps in the identification of Hypoperfused Areas. - Assessment of Collateral Circulation. - Provides guidance for Reperfusion Therapy. - Non-Invasive imaging modality, making it safer for repeated imaging.
Disadvantages	- Prone to false positives, especially for small vessel disease or artifacts. - Less effective in identifying very small or lacunar infarcts.	- Can be influenced by various factors such as age and the presence of other pathologies - Requires careful alignment with DWI to avoid misinterpretation. - Requires careful alignment with DWI to avoid misinterpretation.	- Requires contrast agents usage, which can be contraindicated in patients with kidney issues. - Interpretation can be complex and requires specialized expertise. - Time-consuming compared to other imaging modalities.	- Can overestimate the degree of stenosis or miss very small vessel occlusions. - May be less effective in patients with irregular heartbeats, which can cause artifacts. - Limited spatial resolution compared to conventional angiography.	- Less sensitive to very early ischemic changes compared to DWI. - Can be affected by patient movement, leading to artifacts. - Takes longer to acquire than some other MRI sequences.	- Susceptible to artifacts from metallic objects. - Can produce susceptibility artifacts, particularly near air-tissue interfaces. - Limited in distinguishing between different types of hemorrhages.	- Susceptible to artifacts from metallic objects and air-tissue interfaces. - Longer acquisition time compared to standard sequences. - Interpretation can be challenging due to the complexity of signal changes.	- Susceptibility to motion Artifacts. - Requires Gadolinium-Based Contrast Agents, which can be contraindicated in patients with severe renal impairment. - Limited Quantitative Information on CBV and CBF. - Requires expertise for accurate interpretation.
CT modalities	**NCCT**	**CBF**	**CBV**	**MTT**	**Tmax**	**CTA**	**ASPECTS**	
Advantages	- Rapid, widely available, and easy to perform. - Excellent for detecting acute hemorrhage. - Helps in the initial assessment of stroke type (ischemic vs. hemorrhagic).	- Measures the amount of blood passing through a given volume of brain tissue per unit of time. - Identifies areas of reduced perfusion, critical for determining the ischemic penumbra. - Helps differentiate salvageable tissue from the infarct core.	- Measures the total volume of blood within a given volume of brain tissue. - Helps identify the infarct core, as severely reduced CBV indicates irreversible damage. - Assesses the extent of ischemic damage and potential for tissue recovery.	- Measures the average time it takes for blood to pass through a given brain region. - Prolonged MTT indicates areas of hypoperfusion, useful for identifying at-risk tissue. - Helps in assessing the severity and extent of perfusion deficits.	- Measures the time it takes for the contrast to reach its peak concentration in brain tissue. - Tmax > 6 seconds is often used to identify tissue at risk of infarction. - Provides additional information on the extent of perfusion deficits.	- Non-invasive and rapid assessment of large vessel occlusions and stenosis. - Aids in planning interventional procedures like thrombectomy. - Detects atherosclerotic disease, dissections, and other vascular pathologies.	- Provides a standardized method for evaluating early ischemic changes. - Predicts functional outcomes and guides treatment decisions for reperfusion therapies. - Simple and quick to apply, facilitating rapid clinical decision-making.	-
Disadvantages	- Limited sensitivity for early ischemic changes. - Can miss small infarcts and subtle parenchymal changes. - Provides less detailed information on tissue viability compared to advanced modalities.	- Susceptible to motion artifacts. - Requires precise arterial input function (AIF) determination for accuracy.	- Limited in distinguishing between benign oligemia and ischemic penumbra.	- Interpretation can be complicated by variations in collateral blood flow.	- Susceptible to delays caused by collateral circulation, which can affect accuracy.	- Requires iodinated contrast agents, which can be contraindicated in patients with renal insufficiency. - Artifacts can occur, especially in the posterior fossa and with heavy calcifications.	- Subjective and dependent on the experience of the radiologist. - Less sensitive to small or subtle infarcts. - Can be affected by pre-existing brain abnormalities, such as old infarcts or white matter disease.	-

treatment. On the other hand, fewer studies developed stroke diagnostic approaches ($n = 3$). Such approaches focus on stroke detection and prediction of risk of stroke. In particular, two studies classified patients as stroke versus non-stroke [8], [56].

IV. DISCUSSION

In this literature review, we investigated multimodal machine learning methods presented in recent state-of-the-art work for stroke prognosis and diagnosis. By following PRISMA guidelines for literature search and selection, we specifically analyzed the characteristics of 33 research articles published between January 2015 and March 2023. We discuss the advantages and disadvantages of the imaging modalities in Table VIII, of other clinical modalities in Table IX, and of the most commonly-used models in Table X. In the next few sections, we discuss pertinent

TABLE IX
SUMMARY OF ADVANTAGES AND DISADVANTAGES OF CLINICAL AND TIME-SERIES DATA

Modalities		
	Clinical data (EHR)	**Time series (e.g., EEG, ECG)**
Advantages	- Suitable for predictive modeling. - Enables continuous monitoring of patient health.	- Provide real-time monitoring of patient's status. - Non-invasive assessment of brain and heart conditions. - Widespread availability.
Disadvantages	- Often heterogeneous, incomplete, and stored in separate systems. - Requires extensive preprocessing.	- Susceptible to artifacts and noise. - Limited spatial resolution. - Interpretation requires expertise.

TABLE X
SUMMARY OF THE ADVANTAGES AND DISADVANTAGES OF THE MOST COMMONLY-USED MACHINE LEARNING MODELS

Methods	Tree-based models	SVM models	Recurrent models	CNN models	U-Net models
Advantages	- Easy to understand and interpret. - Provide a visual illustration of the decisions made. - Fast and scalable, able to deal with large and high-dimensional datasets. - Robust and flexible, able to handle missing values, outliers and nonlinear relationships.	- Work better when there is a clear separation between classes. - Effective in high dimensional spaces than lower dimensional space. - More efficient when the number of dimensions is greater than the number of samples. - Work well with small to medium-sized datasets.	- Efficient in modelling sequential data. - Able to capture the dependencies and relationships within sequences due to its internal memory. - Able to process sequences of variable lengths. - Computationally efficient due to parameter sharing.	- Able to learn from raw imaging data automatically, without the need for feature engineering. - Able to exploit the spatial and hierarchical structure of the images. - Able to provide highly accurate results matching human performance. - Less sensitive to noisy data.	- Flexible and can be used for any image segmentation task. - Provide highly accurate results with proper training settings. - Known to achieve good performance in biomedical applications. - Omit the use of fully connected layers.
Disadvantages	- Susceptible to overfitting and underfitting, depending on the tree size. - Susceptible to significant changes with respect to small variations in the data. - Require extensive regularization during training.	- Require long training time, especially with large datasets. - Difficult to tune as they have multiple hyperparameters. - Suffer from limited interpretability. - Sensitive to noisy data.	- Suffer from vanishing or exploding gradients. - Perform poorly with long sequences due to its limited internal memory. - Suffer from the lack of global context which makes them biased towards most recent data.	- Require extensive computational and memory resources. - Require large amounts of training data. - Prone to overfitting.	- Require extensive computational and memory resources. - Require large amounts of training data.

considerations, challenges, and future directions surrounding the development and application of multimodal learning for stroke prognosis and diagnosis.

A. Advanced Multimodal Deep Learning Frameworks

Based on the included studies, we note that statistical machine learning methods require resource-extensive feature engineering pipelines. Deep neural networks, on the other hand, are able to process raw data end-to-end, including imaging, time-series, and clinical data. Deep neural networks enable avenues for automated data processing, as well as modeling tasks beyond simple fusion of features, since multimodal data representations can be co-learned from scratch. One promising methodological direction is to propose new neural network architectures through the design of appropriate encoders for stroke-related modalities, or a unified architecture that is able to process diverse modalities, such as transformers [86]. Incorporating attention modules can potentially improve inter-modal intra-modal interplay, leading to performance improvements. It can also improve model interpretability by learning a relative weighing of the different modalities during training and by indicating their relative importance at inference.

Deep neural networks can also be incorporated in frameworks that leverage multimodal learning paradigms beyond information fusion, such as alignment and co-learning. The former aims at discovering agreement between components of various modalities, while co-learning exploits complementary information between modalities to supplement low-resource modalities with rich-resource modalities [24]. Lastly, the adoption of causal discovery as part of the multimodal learning framework has the potential to unveil causal relationships that govern stroke-related outcomes. Integrating such approaches in the context of stroke prognosis and diagnosis can pave the way for novel methodology and research questions.

B. Computational and Space Complexity

A relevant aspect worth of consideration is the computational and space complexity of the developed multimodal approaches for stroke prognosis and diagnosis. Computational complexity in this context refers to the amount of computing time required to train each model, whereas space complexity refers to the amount of memory or storage needed. Deep learning models typically involve high computational complexity since they have a large number of parameters and layers, often surpassing tens of millions of parameters. In general, the training time is proportional to the number of layers and the size and dimensionality of the training datasets [87]. In terms of space complexity, deep neural networks require significant storage space to allocate the large number of parameters (weights and biases) of each layer. On the other hand, statistical machine learning methods such as SVM and Random Forest typically have lower computational complexity. For instance, SVM has an approximated training complexity of $O(n^3)$ while Random Forest of $O(t \cdot n \cdot log(n))$, where t refers to the number of trees [88], [89] and n is the number of training samples. The space complexity of statistical models is usually much lower than those of neural networks.

With multimodality, space and computational complexity increase, especially for deep neural network architectures. Understanding the computational and space complexity of multimodal models is paramount for real-world applications, as these factors directly impact the feasibility of deploying such models in clinical settings. Complex models, which require high computational and space resources, might have limited adoption in real-world settings, especially in resource-constrained environments.

C. Robust Multimodal Models

Multimodal data collected in real-world settings is characterized by sparsity, due to missing modalities, and scarcity, due to the limited availability of labeled datasets. The latter is highlighted by the size of the datasets used in the included studies, which are significantly smaller than those in other multimodal applications, such as image captioning, emotion recognition, and question answering. These two challenges motivate the development of multimodal learning frameworks that are robust to data sparsity and scarcity. To deal with missing modalities, translation may be used as a pre-training step to learn how to map between modalities, or alternatively generate one modality from another.

Generative models may also be used to generate synthetic data that imitates the distribution of real-world data, such as variational auto-encoders or generative adversarial networks. The synthetic data can then be used to further train multimodal neural networks and improve their robustness in terms of model performance. Such approaches can be of great value especially when dealing with class imbalance, such as due to the lower prevalence of hemorrhagic stroke compared to ischemic stroke. For example, Google released the EHRSafe framework, which enables the generation of synthetic EHR data with high-fidelity and privacy-preserving power [90]. Leveraging such capabilities allows for diverse data modality synthesis and alleviates data scarcity.

Other approaches to tackle robustness include model pre-training strategies that leverage large-scale unlabeled data. For example, self-supervised pretraining enables the model to learn

semantic information prior to fine-tuning on a downstream classification task, reducing the need for annotated datasets.

D. Federated Learning

One solution for the creation of larger multimodal stroke datasets is data sharing across institutions. Such efforts can even lead to more diverse datasets to facilitate the development of high-impact research. A good example is the BIGPICTURE project,[1] which aims to build the largest database of pathology images for machine learning research. As for stroke, the second leading cause of mortality in the world, similar initiatives would be of great value and may have a significant impact not only on the lives of targeted individuals but on society overall.

Yet, data sharing is often challenging due to the sensitivity and nature of clinical data, hence being subject to ethical considerations and strict regulatory requirements. Federated learning in the context of stroke prognosis and diagnosis from a multimodal perspective is one promising area of future work. In the federated learning setting, collaborative models are trained and shared across several clients without the need for data sharing. Such solutions can indirectly improve the issue of data accessibility, while preserving privacy.

E. Clinical Utility of Multimodal Learning

Another major consideration for multimodal learning for stroke prognosis and diagnosis is to ensure that the proposed solutions have high clinical utility. The increase in the number of modalities required for computing a prediction may incur higher resources in real-world settings upon deployment. For example, despite the benefits of MRI imaging, including high sensitivity to small-sized infarcts, MRI scanning facilities are not always readily available in emergency settings, and the acquisition process is time-consuming and may not be suitable for all patients, such as patients with metallic implants. For those reasons, MRI may be clinically less preferred compared to cheaper alternatives, like CT scans. Such constraints limit the scalability and clinical utility of approaches that rely on expensive input modalities, although they may result in more effective diagnosis or prognosis. This further emphasizes the need to design multimodal learning frameworks with appropriate considerations and assumptions that reflect the nature of data collection in real-world settings.

Exploring the use of less-explored routine modalities can contribute to model performance improvement and widen the depth of insights derived from multimodal learning signals. For example, Optical Coherence Tomography (OCT) have demonstrated promising results as a diagnostic tool for Alzheimer's disease [26] while fundus photography has shown a significant role in diagnosing cardiovascular diseases [91]. Other examples include data modalities collected by wearable devices, such as gait patterns, motor function, and daily activities, which can lead to a better understanding of the patient's prognosis.

F. Unified Multi-Task Multimodal Learning Systems

One limitation of existing multimodal learning frameworks is that they are tailored to perform a single prediction task. For future work, we propose the development of a unified multi-task multimodal learning system, in which for a given patient, it is able to predict their future risk of stroke, detect the presence of stroke and its type, and predict prognosis. Finally, it is worth noting that such advancements require effective multi-disciplinary collaboration between biomedical engineering, machine learning and clinical expertise, to translate such advancements into real-world positive impact.

[1]https://www.bigpicture.eu/

V. Conclusion

We conducted a thorough literature review on multimodal machine learning for stroke prognosis and diagnosis. Our findings reveal that multimodal machine learning models provide high accuracy for stroke prognosis and diagnosis as compared to unimodal models. There is also strong high-quality evidence supporting their use in clinical practice, indicating that these models can be effectively utilized in practice for clinical decision-making. The main strength of our study is that we provide a comprehensive overview of primary datasets, data modalities, and learning methodologies pertaining to stroke prognosis and diagnosis.Our aim is to shed light on future research directions to advance the developments in the field of stroke prognosis and diagnosis. Potential research directions include but not limited to building sophisticated and robust multimodal deep learning algorithms, federated learning approaches employment, maintaining the balance between clinical utility and the utilized data modalities of the newly developed methods, and multimodal multi-tasking. This would ultimately contribute to improving health outcomes of patients with stroke.

In terms of limitations, we confined the selection process to top-tier venues, which led to the rejection of some relevant work published in lower-ranked venues. Second, our data extraction process focused on understanding the methodological components of the included papers, while other elements such as the demographics of the patient cohorts included within the datasets were not investigated. Lastly, the interpretation of the presented results relies on the accuracy of manually extracted information, which may entail unintentional errors. We tried to minimize the impact the aforementioned limitations to improve the quality of the systematic review. To the best of our knowledge, this is the first systematic review work that discusses stroke prognosis and diagnosis from the perspective of multimodal machine learning. Significant efforts are needed to advance existing methodology and meet the full potential of multimodal learning for stroke prognosis and diagnosis.

Acknowledgment

The authors would like to thank Sarah Pardo and Aadim Nepal for their helpful feedback.

References

[1] V. L. Feigin et al., "Global, regional, and national burden of stroke and its risk factors, 1990–2019: A. systematic analysis for the global burden of disease study 2019," *Lancet Neurol.*, vol. 20, pp. 795–820, 2021.

[2] M. Lee, Y.-L. Wu, and B. Ovbiagele, "Trends in incident and recurrent rates of first-ever ischemic stroke in Taiwan between 2000 and 2011," *J. Stroke*, vol. 18, no. 1, pp. 60–65, 2016.

[3] V. L. Feigin et al., "World Stroke Organization (WSO): Global stroke fact sheet 2022," *Int. J. Stroke*, vol. 17, pp. 18–29, 2022.

[4] M. Katan and A. Luft, "Global burden of stroke," *Seminars Neurol.*, vol. 38, 2018, pp. 208–211.

[5] A. P. Coupland et al., "The definition of stroke," *J. Roy. Soc. Med.*, vol. 110, pp. 9–12, 2017.

[6] R. von Kummer, "Time is brain: Fact or fiction," *Stroke*, vol. 50, pp. 552–553, 2019.

[7] B.-O. Bat-Erdene and J. L. Saver, "Automatic acute stroke symptom detection and emergency medical systems alerting by mobile health technologies: A review," *J. Stroke Cerebrovascular Dis.*, vol. 30, 2021, Art. no. 105826.

[8] T. Cai et al., "Deepstroke: An efficient stroke screening framework for emergency rooms with multimodal adversarial deep learning," *Med. Image Anal.*, vol. 80, 2022, Art. no. 102522.

[9] M. Saini et al., "Silent stroke: Not listened to rather than silent," *Stroke*, vol. 43, no. 11, pp. 3102–3104, 2012.

[10] C. Ellis, J. Barley, and A. Grubaugh, "Poststroke knowledge and symptom awareness: A global issue for secondary stroke prevention," *Cerebrovasc. Dis.*, vol. 35, no. 6, pp. 572–581, 2013.

[11] S. C. Johnston, "Transient ischemic attack," *N. England J. Med.*, vol. 347, no. 21, pp. 1687–1692, 2002.

[12] J. L. Wilson et al., "Improving the assessment of outcomes in stroke: Use of a structured interview to assign grades on the modified Rankin scale," *Stroke*, vol. 33, no. 9, pp. 2243–2246, 2002.

[13] T. J. Quinn et al., "Reliability of the modified Rankin scale: A systematic review," *Stroke*, vol. 40, no. 10, pp. 3393–3395, 2009.

[14] M. Goyal, J. M. Ospel, M. Kappelhof, and A. Ganesh, "Challenges of outcome prediction for acute stroke treatment decisions," *Stroke*, vol. 52, no. 5, pp. 1921–1928, 2021.

[15] I. R. Chavva et al., "Deep learning applications for acute stroke management," *Ann. Neurol.*, vol. 92, pp. 574–587, 2022.

[16] A. Subudhi et al., "Application of machine learning techniques for characterization of ischemic stroke with MRI images: A review," *Diagnostics*, vol. 12, no. 10, 2022, Art. no. 2535.

[17] Y. Zhang, S. Liu, C. Li, and J. Wang, "Application of deep learning method on ischemic stroke lesion segmentation," *J. Shanghai Jiaotong Univ. (Sci.)*, vol. 27, no. 1, pp. 99–111, 2022.

[18] J. Soun et al., "Artificial intelligence and acute stroke imaging," *Amer. J. Neuroradiol.*, vol. 42, no. 1, pp. 2–11, 2021.

[19] K. Mouridsen, P. Thurner, and G. Zaharchuk, "Artificial intelligence applications in stroke," *Stroke*, vol. 51, no. 8, pp. 2573–2579, 2020.

[20] R. Karthik, R. Menaka, A. Johnson, and S. Anand, "Neuroimaging and deep learning for brain stroke detection-a review of recent advancements and future prospects," *Comput. Methods Programs Biomed.*, vol. 197, 2020, Art. no. 105728.

[21] N. M. Murray et al., "Artificial intelligence to diagnose ischemic stroke and identify large vessel occlusions: A systematic review," *J. Neurointerv. Surg.*, vol. 12, no. 2, pp. 156–164, 2020.

[22] N. A. Shlobin et al., "Artificial intelligence for large vessel occlusion stroke: A systematic review," *World Neurosurgery*, vol. 159, pp. 207–220.e1, 2021.

[23] M. S. Sirsat, E. Fermé, and J. Câmara, "Machine learning for brain stroke: A review," *J. Stroke Cerebrovascular Dis.*, vol. 29, no. 10, 2020, Art. no. 105162.

[24] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, Feb. 2019.

[25] R. Yan et al., "Integration of multimodal data for breast cancer classification using a hybrid deep learning method," in *IEEE Int. Conf. Intell. Comput.*, 2019, pp. 460–469.

[26] C. E. Wisely et al., "Convolutional neural network to identify symptomatic Alzheimer's disease using multimodal retinal imaging," *Brit. J. Ophthalmol.*, vol. 106, no. 3, pp. 388–395, 2022.

[27] F. Mohsen, H. Ali, N. E. Hajj, and Z. Shah, "Artificial intelligence-based methods for fusion of electronic health records and imaging data," *Sci. Rep.*, vol. 12, no. 1, pp. 1–16, 2022.

[28] F. Behrad and M. S. Abadeh, "An overview of deep learning methods for multimodal medical data mining," *Expert Syst. Appl.*, vol. 200, 2022, Art. no. 117006.

[29] S.-C. Huang et al., "Fusion of medical imaging and electronic health records using deep learning: A systematic review and implementation guidelines," *NPJ Digit. Med.*, vol. 3, 2020, Art. no. 136.

[30] M. A. Azam et al., "A review on multimodal medical image fusion: Compendious analysis of medical modalities, multimodal databases, fusion techniques and quality metrics," *Comput. Biol. Med.*, vol. 144, 2022, Art. no. 105253.

[31] Y.-D. Zhang et al., "Advances in multimodal data fusion in neuroimaging: Overview, challenges, and novel orientation," *Inf. Fusion*, vol. 64, pp. 149–187, 2020.

[32] Z. Liu et al., "Machine learning for multimodal electronic health records-based research: Challenges and perspectives," 2021, *arXiv:2111.04898*.

[33] G. Muhammad et al., "A comprehensive survey on multimodal medical signals fusion for smart healthcare systems," *Inf. Fusion*, vol. 76, pp. 355–375, 2021.

[34] S. Shurrab et al., "Multi-modal machine learning for stroke prognosis and diagnosis: A systematic review" 2023. [Online]. Available: https://www.crd.york.ac.uk/prospero/display_record.php?ID=CRD42023390730

[35] M. J. Page et al., "The PRISMA 2020 statement: An updated guideline for reporting systematic reviews," *Syst. Rev.*, vol. 10, no. 1, pp. 1–11, 2021.

[36] P. F. Whiting et al., "Quadas-2: A revised tool for the quality assessment of diagnostic accuracy studies," *Ann. Intern. Med.*, vol. 155, no. 8, pp. 529–536, 2011.

[37] J. Lee et al., "Quapas: An adaptation of the quadas-2 tool to assess prognostic accuracy studies," *Ann. Intern. Med.*, vol. 175, no. 7, pp. 1010–1018, 2022.

[38] O. Maier, B. Menze, and M. Reyes, "Isles challenge 2015: Ischemic stroke lesion segmentation," 2015. Accessed: Jul. 7, 2023. [Online]. Available: http://www.isles-challenge.org/ISLES2015/

[39] K. Egger, M. Maier, O. Reyes, and R. Wiest, "Isles challenge 2016: Ischemic stroke lesion segmentation," 2016, last accessed on Jul. 7, 2023. [Online]. Available: http://www.isles-challenge.org/ISLES2016/

[40] H. Arsany, M. Reyes, R. Wiest, and S. Winzeck, "Isles challenge 2017: Ischemic stroke lesion segmentation," 2017. Accessed: Jul. 7, 2023. [Online]. Available: http://www.isles-challenge.org/ISLES2017/

[41] H. Arsany et al., "Isles challenge 2018: Ischemic stroke lesion segmentation," 2018. Accessed: Jul. 7, 2023. [Online]. Available: http://www.isles-challenge.org/ISLES2018/

[42] P. S. Fransen et al., "MR CLEAN, a multicenter randomized clinical trial of endovascular treatment for acute ischemic stroke in The Netherlands: Study protocol for a randomized controlled trial," *Trials*, vol. 15, no. 1, pp. 1–11, 2014.

[43] H. Menon, "Measuring collaterals with multi-phase ct angiography in patients with ischemic stroke (prove-it)," 2017. Accessed: Jul. 12, 2023. [Online]. Available: https://classic.clinicaltrials.gov/ct2/show/NCT02184936

[44] A. M. Demchuk et al., "Endovascular treatment for small core and anterior circulation proximal occlusion with emphasis on minimizing ct to recanalization times (escape) trial: Methodology," *Int. J. Stroke*, vol. 10, pp. 429–438, 2015.

[45] H. C. d. Lyon, "Cohort of ischemic stroke patients (hibiscusstroke)," 2020. Accessed: Jul. 11, 2023. [Online]. Available: https://classic.clinicaltrials.gov/ct2/show/NCT03149705

[46] B. Hotter et al., "Prospective study on the mismatch concept in acute stroke patients within the first 24 h after symptom onset-1000plus study," *BMC Neurol.*, vol. 9, 2009, Art. no. 60.

[47] M. Goyal et al., "Endovascular thrombectomy after large-vessel ischaemic stroke: A meta-analysis of individual patient data from five randomised trials," *Lancet*, vol. 387, pp. 1723–1731, 2016.

[48] C. Wang et al., "Deep learning-based identification of acute ischemic core and deficit from non-contrast ct and cta," *J. Cereb. Blood Flow Metab*, vol. 41, no. 11, pp. 3028–3038, 2021.

[49] A. Clérigues et al., "Acute ischemic stroke lesion core segmentation in CT perfusion images using fully convolutional neural networks," *Comput. Biol. Med.*, vol. 115, 2019, Art. no. 103487.

[50] Z. Liu et al., "Towards clinical diagnosis: Automated stroke lesion segmentation on multi-spectral MR image using convolutional neural network," *IEEE Access*, vol. 6, pp. 57006–57016, 2018.

[51] J. Rubin and S. M. Abulnaga, "CT-to-MR conditional generative adversarial networks for ischemic stroke lesion segmentation," *Proc. IEEE Int. Conf. Healthcare Inform.*, 2019, pp. 1–7.

[52] G. Brugnara et al., "Multimodal predictive modeling of endovascular treatment outcome for acute ischemic stroke using machine-learning," *Stroke*, vol. 51-12, pp. 3541–3551, 2020.

[53] E. Zihni et al., "Multimodal fusion strategies for outcome prediction in stroke," in *Proc. 13th Int. Joint Conf. Biomed. Eng. Syst. Technol.*, 2020, pp. 421–428.

[54] Z. A. Samak, P. Clatworthy, and M. Mirmehdi, "Prediction of thrombectomy functional outcomes using multimodal data," in *Proc. Annu. Conf. Med. Imag. Understanding Anal.*, 2020, pp. 267–279.

[55] S. Bacchi et al., "Deep learning in the prediction of ischaemic stroke thrombolysis functional outcomes: A pilot study," *Academic Radio.*, vol. 27, pp. e19–e23, 2020.

[56] J. Yu et al., "Ai-based stroke disease prediction system using ECG and PPG bio-signals," *IEEE Access*, vol. 10, pp. 43623–43638, 2022.

[57] A. M. Chiarelli et al., "Electroencephalography-derived prognosis of functional recovery in acute stroke through machine learning approaches," *Int. J. Neural Syst.*, vol. 30, no. 12, 2020, Art. no. 2050067.
[58] M. Roder, G. H. Rosa, and J. P. O. Papa, "Enhancing shallow neural networks through fourier-based information fusion for stroke classification," in *Proc. IEEE 34th SIBGRAPI Conf. Graph., Patterns Images*, 2021, pp. 378–385.
[59] H. Yu et al., "Prognosis of ischemic stroke predicted by machine learning based on multi-modal MRI radiomics," *Front Psychiatry*, vol. 13, 2022, Art. no. 1105496.
[60] D. Rajashekar et al., "Prediction of clinical outcomes in acute ischaemic stroke patients: A comparative study," *Front Neurol*, vol. 12, 2021, Art. no. 663899.
[61] H. Kuang et al., "Computed tomography perfusion–based machine learning model better predicts follow-up infarction in patients with acute ischemic stroke," *Stroke*, vol. 52, pp. 223–231, 2021.
[62] G. X. Chen et al., "RFDCR: Automated brain lesion segmentation using cascaded random forests with dense conditional random fields," *Neuroimage*, vol. 211, 2020, Art. no. 116620.
[63] R. Karthik, R. Menaka, M. Hariharan, and D. Won, "Ischemic lesion segmentation using ensemble of multi-scale region aligned CNN," *Comput. Methods Prog. Biomed.*, vol. 200, 2021, Art. no. 105831.
[64] L. Liu et al., "Deep convolutional neural network for automatically segmenting acute ischemic stroke lesion in multi-modality MRI," *Neural Comp. Appl.*, vol. 32, pp. 6545–6558, 2020.
[65] Y. Xie et al., "Use of gradient boosting machine learning to predict patient outcome in acute ischemic stroke on the basis of imaging, demographic, and clinical information," *AJR Amer. J. Roentgenol.*, vol. 212, no. 1, pp. 44–51, 2019.
[66] W. Qiu et al., "Radiomics-based intracranial thrombus features on CT and CTA predict recanalization with intravenous alteplase in patients with acute ischemic stroke," *AJNR Amer. J. Neuroradiol*, vol. 40, pp. 39–44, 2019.
[67] H. Cui, X. Wang, Y. Bian, S. Song, and D. D. Feng, "Ischemic stroke clinical outcome prediction based on image signature selection from multimodality data," *Proc. 40th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2018, pp. 722–725.
[68] A. Pinto et al., "Stroke lesion outcome prediction based on MRI imaging combined with clinical information," *Front. Neurol.*, vol. 9, 2018, Art. no. 1060.
[69] R. McKinley et al., "Fully automated stroke tissue estimation using random forest classifiers (FASTER)," *J. Cereb. Blood Flow Metab.*, vol. 37, no. 8, pp. 2728–2741, 2017.
[70] M. Livne et al., "Boosted tree model reforms multimodal magnetic resonance imaging infarct prediction in acute stroke," *Stroke*, vol. 49, pp. 912–918, 2018.
[71] T. Y. Shi, H. Y. Jiang, and B. Zheng, "C ^{2} MA-Net: Cross-modal cross-attention network for acute ischemic stroke lesion segmentation based on CT perfusion scans," *IEEE Trans. Biomed. Eng.*, vol. 69, no. 1, pp. 108–118, Jan. 2022.
[72] A. Pinto, J. Amorim, A. Hakim, V. Alves, M. Reyes, and C. A. Silva, "Prediction of stroke lesion at 90-day follow-up by fusing raw DSC-MRI with parametric maps using deep learning," *IEEE Access*, vol. 9, pp. 26260–26270, 2021.
[73] C. Wang et al., "Data fusion framework for the prediction of early hematoma expansion based on CNN," in *Proc. IEEE 18th Int. Symp. Biomed. Imag.*, 2021, pp. 169–173.
[74] D. Robben et al., "Prediction of final infarct volume from native ct perfusion and treatment parameters using deep learning," *Med. Image Anal.*, vol. 59, 2020, Art. no. 101589.
[75] J. Amorim, A. Pinto, S. Pereira, and C. A. Silva, "Segmentation squeeze-and-excitation blocks in stroke lesion outcome prediction," in *Proc. IEEE 6th Portuguese Meeting Bioeng.*, 2019, pp. 1–4.
[76] Y. Xue et al., "A multi-path 2.5 dimensional convolutional neural network system for segmenting stroke lesions in brain MRI images," *Neuroimage Clin.*, vol. 25, 2020, Art. no. 102118.
[77] A. Pinto et al., "Enhancing clinical MRI perfusion maps with data-driven maps of complementary nature for lesion outcome prediction," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2018, pp. 107–115.
[78] N. Hatami et al., "CNN-LSTM based multimodal mri and clinical data fusion for predicting functional outcome in stroke patients," *44th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2022, pp. 3430–3434.
[79] M. Soltanpour, R. Greiner, P. Boulanger, and B. Buck, "Ischemic stroke lesion prediction in ct perfusion scans using multiple parallel u-nets following by a pixel-level classifier," in *Proc. IEEE 19th Int. Conf. Bioinf. Bioeng.*, 2019, pp. 957–963.
[80] R. McKinley, L. Häni, R. Wiest, and M. Reyes, "Segmenting the ischemic penumbra: A decision forest approach with automatic threshold finding," in *Proc. Brainlesion: Glioma, Multiple Scler., Stroke Traumatic Brain Injuries: 1st Int. Workshop*, 2016, pp. 275–283.
[81] C. Rorden, H.-O. Karnath, and L. Bonilha, "Improving lesion-symptom mapping," *J. Cogn. Neurosci.*, vol. 19, pp. 1081–1088, 2007.
[82] M. Robnik-Šikonja et al., "An adaptation of relief for attribute estimation in regression," in *Mach. Learn.: Proc. the 14th Int. Conf.*, 1997, vol. 5, pp. 296–304.
[83] R. Karthik and R. Menaka, "Computer-aided detection and characterization of stroke lesion–a short review on the current state-of-the art methods," *Imag. Sci. J.*, vol. 66, no. 1, pp. 1–22, 2018.
[84] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," in *Proc. SSST-8, 8th Workshop Syntax, Semantics Struct. Stat. Transl.*, 2014
[85] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, Aug. 2013.
[86] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations*, 2021.
[87] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT press, 2016.
[88] Y. Hou and X. F. Zheng, "SVM based MLP neural network algorithm and application in intrusion detection," in *Proc. Artif. Intell. Computat. Intell.: 3rd Int. Conf.*, Taiyuan, China, 2011, pp. 340–345.
[89] D. Vos and S. Verwer, "Efficient training of robust decision trees against adversarial examples," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 10586–10595.
[90] J. Yoon et al., "EHR-Safe: Generating high-fidelity and privacy-preserving synthetic electronic health records," *NPJ Digit. Med.*, vol. 6, 2022, Art. no. 141.
[91] R. Poplin et al., "Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning," *Nature Biomed. Eng.*, vol. 2, no. 3, pp. 158–164, 2018.