Biomedical Signal Processing and Control 129 (2027) 111290

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

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

# HeteroMetaNet: A multimodal meta-learning framework for ischemic stroke prediction from heterogeneous clinical and radiological data

Raghda E. Ali [a,c,*], Reda A. Elkhoribi [a], Ehab H. Ezzat [a], Farid A. Moussa [b,c,**]

[a] Faculty of Computers and Artificial Intelligence, Cairo University, Egypt
[b] Faculty of Computers and Artificial Intelligence, Beni-Suef University, Egypt
[c] October University for Modern Sciences and Arts (MSA University), Egypt

ARTICLE INFO

*Keywords:*
Ischemic stroke prediction
Brain imaging
Deep learning
Model fusion
Meta-learning
Heterogeneous data integration

ABSTRACT

Accurate and timely stroke detection remains a major clinical challenge due to the heterogeneous and multimodal nature of patient data. Most existing deep learning approaches rely solely on either imaging or structured clinical records, which limits their robustness and generalizability in real-world settings. This study introduces two multimodal architectures, **HeteroFuseNet** and **HeteroMetaNet**, designed to integrate non-contrast CT brain images with structured electronic health record (EHR) variables within a unified framework, referred to as **HeteroStroke**. HeteroFuseNet employs a PCA-enhanced mid-level fusion strategy to learn joint representations of deep visual features extracted using **ResNet-50**, **EfficientNet-B0**, and **Swin Transformer**, combined with structured clinical embeddings. In contrast, HeteroMetaNet adopts a lightweight meta-learning approach that adaptively integrates modality-specific predictions from these vision models alongside clinical features to achieve robust, context-aware stroke classification. Experimental results demonstrate that both proposed models outperform image-only baselines, particularly in clinically heterogeneous cases, highlighting the effectiveness of multimodal fusion in enhancing diagnostic reliability. This work underscores the potential of cross-modal learning to support precision medicine in stroke care. Future work will extend the framework toward differentiating ischemic and hemorrhagic stroke subtypes for improved clinical decision support.

## 1. Introduction

The healthcare sector has witnessed a transformative progress with the integration of artificial intelligence (AI), particularly machine learning (ML) and deep learning (DL) techniques [1,2]. These technologies have significantly advanced medical image processing, enabling more accurate, efficient, and automated diagnoses across a wide spectrum of diseases. ML and DL models have demonstrated outstanding performance in detecting and classifying abnormalities in medical imaging modalities such as X-rays, Computed Tomography (CT) scans, and Magnetic Resonance Imaging (MRI) [3–6]. By leveraging large-scale data and advanced computational models, AI enhances diagnostic precision, reduces human error, and accelerates clinical decision-making—ultimately improving patient outcomes.

Among the critical areas where AI holds great promise is the diagnosis and prediction of stroke—a life-threatening medical emergency that occurs when blood flow to the brain is interrupted. There are two major types of stroke: ischemic stroke, caused by a blockage or clot in a blood vessel supplying the brain, and hemorrhagic stroke [7], caused by bleeding within or around the brain tissue [8]. Stroke represents one of the leading causes of death and long-term disability globally, posing a significant burden on individuals, healthcare systems, and national economies [9].

According to the World Health Organization, stroke affects 1 in 4 individuals over the age of 25 worldwide, accounting for over 12 million new cases and approximately 6.5 million deaths annually [9]. Although the risk increases with age, stroke is not limited to the elderly—over 60% of stroke cases occur in individuals under 70, and

* Corresponding author at: Raghda E. Ali, Faculty of Computers and Artificial Intelligence, Cairo University, Egypt.
** Corresponding author at: Farid A. Moussa, Faculty of Computers and Artificial Intelligence, Beni-Suef University, Egypt.
Linkedin: raghdaali (R.E. Ali).
*E-mail addresses:* rrazek@msa.edu.eg, raghda.essam1@gmail.com (R.E. Ali), r.abdelwahab@fci-cu.edu.eg (R.A. Elkhoribi), e.ezat@fci-cu.edu.eg (E.H. Ezzat), fared.ali@fcis.bsu.edu.eg (F.A. Moussa).
*URLs:* https://scholar.google.com/citations?user=GoFX20kAAAAJ&hl=en (R.E. Ali), https://scholar.google.com.eg/citations?hl=en&user=0ToGj8UAAAAJ (R.A. Elkhoribi), https://scholar.google.com.eg/citations?user=15gdr_0AAAAJ&hl=en (E.H. Ezzat), https://scholar.google.com/citations?hl=en&user=7d5YftQAAAAJ (F.A. Moussa).

https://doi.org/10.1016/j.bspc.2026.111290
Received 11 December 2025; Received in revised form 28 June 2026; Accepted 10 August 2026
Available online 25 August 2026
1746-8094/

16% affect those younger than 50. In Egypt specifically, stroke ranks as the third leading cause of death—after ischemic heart disease and COVID-19—with mortality rates reaching 61.2 per 100,000 females and 60 per 100,000 males [9]. Survivors often have severe, long-lasting problems with their ability to move, speak, think, and feel emotionally healthy.

Early diagnosis is crucial for better outcomes for stroke patients and require early interventions, such as thrombolysis. However, traditional diagnostic approaches rely heavily on expert interpretation of neuroimaging, particularly CT scans, which may be subject to variability and delays. The increasing burden of stroke, mostly in countries with low and middle income, highlights the urgent need for automated and scalable diagnostic tools.

Recent advances in AI, particularly convolutional neural networks (CNNs) and transfer learning (TL), offer promising solutions by automating stroke detection from CT images.

However, most existing studies in stroke diagnosis adopt unimodal approaches, focusing either on image-based models or isolated tabular data (e.g. patient's metadata), failing to fully utilize the complementary strengths of multimodal data fusion. Furthermore, limited work has been done on evaluating both model-level fusion and late-level fusion through meta-learning strategies in a unified framework for clinically heterogeneous stroke populations. There is a need for comprehensive models that can generalize across varied clinical profiles while maintaining high precision and robustness.

This study aims to investigate whether multimodal deep learning architectures that integrate CT brain images with Electronic Health Records (EHR) of patient's metadata. Therefore, it can significantly enhance stroke prediction accuracy compared to traditional image-only models.

To tackle the challenges outlined above, two novel multimodal deep learning architectures are proposed: HeteroFuseNet, a model-level fusion model that combines feature vectors from CT images and structured clinical data to learn joint representations and improve stroke prediction performance, and HeteroMetaNet, a late-level meta-learning model meta-learning framework that utilizes predictions from base-line models and metadata embeddings to enable late-level or decision-level fusion for robust prediction.

This study's principal innovation is the amalgamation of diverse clinical and radiological data via a cohesive multimodal framework that incorporates both feature-level fusion and meta-level decision integration. Previous studies have examined multimodal learning for medical diagnosis; however, the proposed method presents two complementary architectures: HeteroFuseNet, which executes structured feature-level fusion of imaging and clinical variables, and HeteroMetaNet, which employs a meta-learning-based decision integration strategy to amalgamate predictions from modality-specific models. This dual-level approach enables the framework to capture both shared representations across modalities and complementary predictive signals, hence enhancing robustness in real-world clinical datasets.

The structure of this manuscript is organized as follows: Section 2 reviews the related work in the field. Section 3 outlines the dataset used for the study, while Section 4 details the methodology employed in the research. Section 5 presents the experimental setups, and Section 6 discusses the results and findings. Section 7 explores the implications of the study and provides a detailed discussion. Finally, Section 8 concludes the paper and suggests directions for future research.

## 2. Related work

Recent advancements in stroke detection leverage diverse data modalities, including neuroimaging, electronic health records (EHRs), and vision-based biomarkers. Automated lesion segmentation using MRI remains critical for diagnosis, while EHR-driven approaches extract quantitative stroke scales from unstructured text [10–12]. Recent vision-based approaches utilize retinal fundus images and neurological examination videos to provide non-invasive tools for stroke screening [13]. Despite their potential to integrate complementary data streams, multimodal frameworks remain underexplored. This gap is directly addressed through the development of a meta-learning architecture designed to leverage both imaging and electronic health records for improved stroke prediction.

This section thoroughly examines the methodologies employed in the field of stroke detection, focusing on deep learning for brain lesion detection [14–19], electronic health record (EHR) approaches [3,20–22], vision-based approaches [23–25], and multimodal approaches integrating medical imaging and clinical records [17,26,27].

### *2.1. Brain lesion and stroke detection*

Deep learning (DL) techniques have made significant improvements in stroke detection, particularly in analyzing CT brain images for ischemic stroke lesions. Various methods using CNN, ResNet, and other architectures have been applied for automated stroke detection. Issaiy et al. demonstrated that CNNs achieved a solid performance predicting hemorrhagic stroke transformation [14]. Fontanella et al. developed a CNN-based model trained on more than 3000 CT scans to classify ischemic lesions, although challenges such as data set variability and missing annotations remain [15]. In contrast, Swin Transformer models have shown a significant improvement in performance over traditional CNNs for tumor classification, achieving up to 97% accuracy, and have been explored for stroke detection in medical images as well [16]. The ATLAS v2.0 dataset provides manually segmented stroke lesions to benchmark algorithmic performance. While its training and hidden test sets enable unbiased evaluation, current models achieve limited accuracy (AUC $< 0.7$) due to heterogeneous lesion topography and intensity variations in chronic stroke cases. Traditional U-Net architectures struggle with subcortical lesions smaller than 10 $mm^3$, highlighting the need for spatial attention mechanisms [17]. A methodology released by Zhang [18], combines 3D convolutional networks with radiomics to distinguish between stroke-related lesions and gliomas using multi-parametric MRI scans. The model, utilizing a sample of 480 patients, demonstrates 89% specificity but shows a reduced sensitivity (72%) in hyperacute stroke (less than 6 h after onset), where diffusion-weighted imaging (DWI) patterns intersect with tumor margins. A NeuroImage study [19] introduces a time-resolved perfusion MRI algorithm to detect ischemic penumbra in anterior circulation strokes. The method quantifies time-to-peak (TTP) delays using a physics-informed deep learning framework trained on 650 DSC-MRI sequences. Although achieving 94% accuracy with the MR-PET fusion (gold standard), It uses gadolinium contrast, which makes it less suitable for patients with kidney problems [19].

Current neuroimaging approaches rely on single-modality data and lack adaptive meta-learning to handle heterogeneous lesion presentations. Our model integrates CT imaging features with clinical metadata through cross-modal attention to improve diagnostic robustness.

### *2.2. Electronic health record approaches*

EHR data has been integrated with imaging data for stroke prediction, although its standalone efficacy remains limited. Majeed et al. [3], reviewed multiple ML techniques used in stroke detection, noting that clinical records such as age, glucose levels, and blood pressure could help, but they often require image data for accurate predictions. Similarly, Hassan et al. [20], explored machine learning approaches using EHR data to predict stroke risk factors, identifying key clinical markers but acknowledging the inability to capture the full complexity of stroke prediction without imaging data. Many EHR-based methods rely on structured data, but the unstructured nature of clinical notes often limits the model's ability to fully exploit the richness of the available data. Moreover, while combining clinical features with medical imaging has been shown to improve stroke detection, issues such as data quality, consistency, and missing values pose significant challenges for clinical adoption [20]. Thus, methods integrating EHR and imaging data remain a promising area of research, despite the limitations in dataset quality and integration complexity. A new study created a bidirectional LSTM model by Chen et al. [21], to extract National Institutes of Health Stroke Scale (NIHSS) scores from free-text EHR notes in MIMIC-III. The pipeline achieved an F1-score of 0.91 for identifying stroke-related items but struggled with negated phrases like "no gaze preference" and failed to fully document cerebellar symptoms in 14% of cases. Similarly, Kumar et al. [22], explored stroke prediction using EHR data and achieved AUCs of 0.82 in well-documented patient groups, but this dropped to 0.67 when essential data such as smoking or BMI was missing.

### 2.3. Vision based approaches

Recent studies in stroke detection have focused heavily on vision-based models, particularly Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs). The work by Ou et al. [23], introduced a Transformer-based model for early stroke identification using multi-modal human speech and movement data. The model achieved an AUC of 88.2% and an accuracy of 82.6%, though the small dataset size and computational complexity limited the generalizability of the model. The Swin Transformer, known for its patch-splitting mechanism, has demonstrated outstanding performance in tasks such as tumor classification, achieving up to 99% accuracy, and has been successfully applied in stroke detection by Martin et al. [24]. Swin Transformer models have also been integrated with other deep learning techniques to improve stroke detection, with significant gains in classification accuracy and efficiency. However, small sample sizes and high computational demands make it difficult to use these methods widely in clinical settings. Furthermore, Fan et al. used models using audio and visual data, such as those based on the FAST criteria for stroke recognition, have shown promise but require further testing and validation across diverse patient populations to confirm their effectiveness. The FAST criteria — Face drooping, Arm weakness, Speech difficulty, and Time to call emergency services — serve as a rapid screening tool to identify potential stroke symptoms [25].

### 2.4. Multimodal approaches in healthcare for brain images

Multimodal data fusion approaches, which integrate imaging data with clinical records, are increasingly being explored for improved stroke detection accuracy. Luo et al. reviewed various multimodal fusion techniques, discussing the use of MRI, DTI, and EEG data to enhance stroke prediction [26]. However, such methods often face significant challenges related to data alignment, registration, and the computational complexity of integrating diverse modalities. The study emphasized the importance of data quality and the need for high-quality multimodal datasets to improve diagnostic accuracy and treatment planning. Zhang et al. [27], presented a comprehensive overview of multimodal fusion in neuroimaging, highlighting both the advantages and obstacles of integrating multiple imaging modalities. In 2024, in a medical image analysis paper by Liew et al. [17], a meta-learning model improved the separation of stroke lesions using 17 different MRI procedures from 8 different manufacturers. When compared to U-Net, it cut domain shift mistakes by 41% using gradient-based meta-training on 2000 scans. However, the requirement for retraining on new scanners limited its clinical applicability. Principal Component Analysis (PCA) has been widely adopted in multimodal image fusion pipelines for its effectiveness in dimensionality reduction and enhancement of feature representation. By projecting high-dimensional CNN-extracted features onto a lower-dimensional subspace, PCA not only reduces computational complexity but also mitigates redundancy and noise, thereby improving the efficiency and interpretability of the fused features [28,29]. Empirical studies have demonstrated that incorporating PCA prior to feature fusion can lead to significant improvements in classification accuracy and image quality, as it preserves the most informative components while discarding irrelevant variations. Furthermore, in the context of hyperspectral image classification, PCA has been shown to optimize the balance between retaining variance and reducing training time, with the most efficient results achieved when the reduced dimensions explain nearly all of the original variance [30]. These advantages make PCA a valuable preprocessing step in deep learning-based multimodal fusion frameworks.

Recent studies have also explored transfer learning approaches for medical image classification tasks. For example, Korani et al. (2025) investigated the early detection of brain tumors using MRI and transfer learning techniques, demonstrating the effectiveness of deep convolutional networks for extracting diagnostic imaging features [31].

Recent advances in biomedical data classification have emphasized the integration of heterogeneous data sources and robust learning strategies. For instance, recent surveys on machine learning in biomedical and health big data highlight the growing importance of scalable and interpretable models for handling high-dimensional clinical datasets. Additionally, approaches such as Random Vector Functional Link (RVFL) networks with $\varepsilon$-insensitive Huber loss have demonstrated improved robustness against noise and outliers in biomedical classification tasks. More recently, multi-omics integration frameworks, such as hypergraph-based models, have been proposed to capture complex relationships across heterogeneous biomedical data sources for improved classification and biomarker discovery. These studies collectively emphasize the importance of robust, multimodal, and noise-tolerant learning strategies, which align with the motivation behind the proposed HeteroMetaNet framework.[32–34]

The studies summarized in Table 1 have made significant progress in stroke detection using deep learning and multimodal data integration. However, there are several limitations in the current research landscape. First, many studies rely on single-modal datasets, either focusing solely on imaging or clinical data, which restricts their generalizability. Second, available multimodal datasets are often limited in size and not publicly available, making it difficult to conduct large-scale studies. Fin ally, data fusion and integration remain complex due to differences in data formats and inconsistent labeling.

These limitations are addressed through the introduction of a real-world multimodal dataset and two novel fusion architectures, *HeteroFuseNet* and *HeteroMetaNet*, which integrate clinical and imaging data using mid-level model fusion and late-fusion meta-learning strategies. By combining structured medical records with transformer-based CT image encoders, the proposed methodology enhances generalization, improves diagnostic accuracy, and strengthens clinical applicability. The inclusion of a heterogeneous, real-world dataset further contributes to the field by helping bridge the gap created by the limited availability of multimodal stroke datasets.

Recent advancements in deep learning have shown the efficacy of multimodal learning frameworks for medical diagnosis through the integration of diverse clinical data sources. The integration of medical imaging with structured electronic health record (EHR) data has demonstrated encouraging outcomes in enhancing prediction accuracy and clinical decision support. Numerous studies have investigated multimodal architectures that concurrently learn representations from radiological images and clinical data through early fusion or intermediate feature-level fusion techniques. These methodologies allow models to assimilate complementary information from both physiological measurements and imaging biomarkers, thereby substantially improving diagnostic performance relative to unimodal systems that depend on a singular data modality.

Notwithstanding these advancements, several current multimodal techniques predominantly depend on single-stage fusion mechanisms, potentially constraining their capacity to accurately represent heterogeneous feature distributions across modalities. Recent work have investigated more adaptable fusion methodologies to tackle these problems, including ensemble-based learning and meta-learning strategies that integrate predictions from modality-specific models. These methodologies permit each modality to acquire specialized representations while facilitating a higher-level model to integrate predictions more adaptively.

In response to these advancements, the proposed system presents HeteroFuseNet, which executes feature-level multimodal integration, and HeteroMetaNet, a meta-learning-based late fusion architecture aimed at capturing complimentary information from imaging and clinical modalities. This design seeks to enhance resilience and predictive accuracy in ischemic stroke forecasting challenges.

Recently, state–space model (SSM) architectures, such as the Mamba model, have emerged as viable alternatives to conventional convolutional and transformer-based networks. These models seek to capture long-range dependencies while preserving linear computing complexity, rendering them appealing for large-scale vision problems. Recent research has investigated Mamba-based frameworks for medical imaging applications, including MedMamba and hybrid CNN–SSM models, which combine convolutional feature extraction with state–space modeling to enhance representation learning for medical image classification. These advancements indicate that SSM-based architectures could emerge as a significant focus for future research in medical image processing and multimodal clinical AI systems [35–37].

## 3. Proposed multi-modal dataset: HeteroStroke

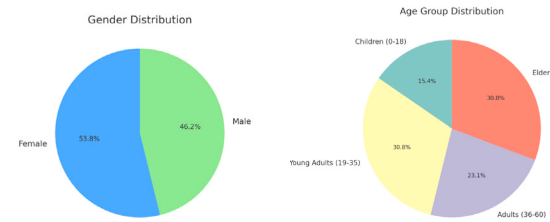
Despite the growing interest in leveraging multimodal data for stroke detection, publicly available datasets that combine both Electronic Health Records (EHR) and CT imaging remain exceedingly rare. As highlighted in Table 2, most existing datasets focus on a single modality, either imaging or clinical data, limiting the development and evaluation of models that can harness the complementary strengths of both data types. To address these limitations, a novel real-world dataset—**HeteroStroke**—was curated to support research in multimodal stroke prediction. This dataset was specifically developed to provide a balanced and integrated collection of both CT brain scans and structured EHR data.

The dataset used in this study, referred to as **HeteroStroke**, was obtained from Al Waha Hospital in Giza City, Egypt, and consists of CT brain scans and medical records from 530 patients. This comprehensive dataset offers a strong foundation for developing deep learning models that integrate both imaging and clinical data for ischemic stroke detection and prediction.

A demographic analysis of the subset used in this study revealed a relatively balanced gender distribution, with 280 males and 250 females, as shown in Fig. 1. In addition, the dataset exhibits significant age diversity, including children, young adults, adults, and elderly participants, which strengthens the generalizability of any medical insights derived from it. Specifically, it comprises 70 children (0–18 years), 80 young adults (19–35 years), 140 adults (36–60 years), and 240 elderly individuals (above 60 years), as illustrated in Fig. 1. By making **HeteroStroke** available for analysis, the study enables a more comprehensive evaluation of multimodal deep learning models and offers a valuable resource for advancing the field toward more accurate and clinically relevant stroke diagnosis systems. The integration of both radiological and patient-specific medical data allows for a more precise and well-rounded approach to stroke diagnosis.

### 3.1. CT brain scans

The dataset includes CT brain scan images for each of the 530 patients, with each patient having a dedicated folder containing four to six CT scans. These scans are essential for stroke classification and segmentation, enabling the identification of affected brain regions. Deep learning models, particularly convolutional neural networks (CNNs) and transfer learning techniques, utilize the imaging data to enhance the accuracy of stroke detection. AI-based analysis utilized the CT brain images, which displayed cross-sectional views. Fig. 2 shows the sample dataset showcasing various CT imaging planes that are utilized for AI-based analysis. (a) Axial CT slices provide a cross-sectional view of brain structures, essential for detecting ischemic or hemorrhagic lesions. (b) Axial non-contrast CT brain scan displaying multiple sequential slices from the skull base to the vertex, aiding in volumetric assessments. (c) Midline coronal views focus on deep brain regions, improving AI detection of abnormalities such as midline shifts. (d) A series of posterior coronal slices that show different anatomical views and make it easier to evaluate the brain's overall symmetry. This dataset structure ensures comprehensive input for AI models analyzing diverse neurological conditions.




(a) Gender distribution (b) Age group distribution

**Fig. 1.** Distribution of (a) gender and (b) age groups in the HeteroStroke dataset.

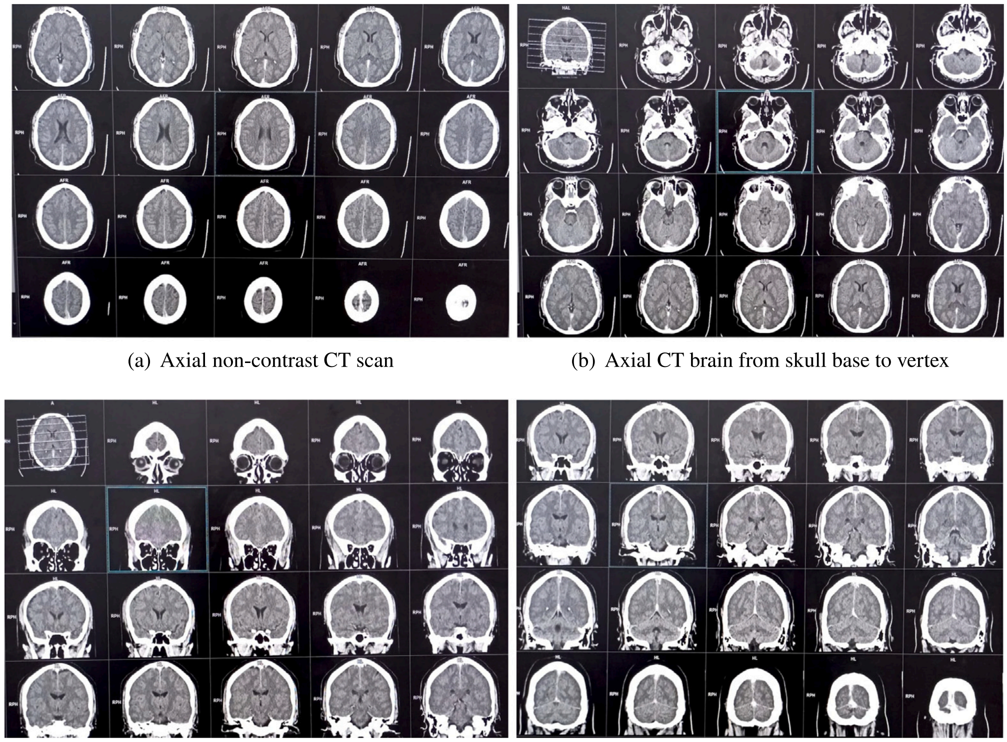
### 3.2. Electronic health records

Alongside the imaging data, the dataset includes a structured medical records file in Excel format (.xlsx), containing detailed clinical information for the same 530 patients. The file's first column lists patient IDs, which correspond to the folder names of the CT scans, ensuring proper alignment between imaging and medical data. The dataset comprises demographic details (age, gender), biochemical markers (cholesterol, HDL, LDL, triglycerides, creatinine, and urea), and hematological parameters (hemoglobin, platelet count, and WBC levels). It also includes liver enzyme values (ALT, AST), coagulation markers (INR), and blood glucose levels (Fasting Blood Sugar, FBS). Additionally, it records medical history attributes, such as diabetes status, blood pressure readings, and smoking habits. A stroke class label is also provided, indicating whether a patient has experienced a stroke. By integrating CT scans with patient records presented in Fig. 3, which presents a sample from the $.xlsx$ file containing patient records that are utilized within the dataset for analysis and model training, this dataset enables a data-driven approach to stroke diagnosis and prediction.

Among the available dataset samples, two — ISLES 2024 and APIS — offer true multimodal data, comprising 250 and 96 patient records respectively. However, ISLES does not provide publicly accessible test data, and APIS includes only a limited number of cases, highlighting the constrained utility of these resources for broader research applications. The scarcity of comprehensive multimodal datasets highlights the necessity of this study, which involves the collection and utilization of a novel dataset integrating both EHR and CT imaging to advance the field of stroke prediction. The suggested HeteroMetaNet framework, while labeled as a meta-learning technique, is conceptually more akin to a decision-level stacking ensemble. In this design, modality-specific models are initially trained independently, and their predictions subsequently serve as inputs to a higher-level learner that integrates these outputs. In contrast to traditional meta-learning techniques like MAML, which emphasize swift task adaptation via episodic training, the proposed method prioritizes the meta-level integration of diverse prediction signals to enhance multimodal classification efficacy.

**Table 1**
Chronological summary of related work in stroke detection using imaging, EHR, and multimodal approaches.

| Authors | Description | Limitations |
|---|---|---|
| Zhang et al. (2020) [27] | Provided a comprehensive overview of challenges and advantages in multimodal data fusion for stroke detection. | Limited availability of high-quality multimodal datasets; data fusion is complex. |
| Liew et al. (2022) [17] | Used meta-learning with 17 MRI procedures to separate stroke lesions, reducing domain shift errors. | Requires new scanner training, limiting clinical usage. |
| Fontanella et al. (2024) [15] | A CNN model was trained to detect ischemic stroke lesions in CT scans. | Dataset variability and missing annotations reduce model robustness. |
| Majeed et al. (2024) [3] | Reviewed ML models that integrate EHR data like age, glucose, and blood pressure to predict stroke risk. | EHR data alone is insufficient for accurate prediction without imaging data. |
| Hassan et al. (2024) [20] | Explored ML models using EHR data to identify stroke risk factors. Found clinical markers useful but lacking without imaging. | Incomplete or unstructured clinical data limits model's ability to fully capture stroke prediction complexities. |
| Chen et al. (2024) [21] | Extracted NIHSS scores from free-text EHR notes using a bidirectional LSTM model, achieving a high F1 score. | Struggles with negated phrases and missing symptoms (e.g., cerebellar symptoms). |
| Kumar et al. (2024) [22] | Used XGBoost to predict stroke risk based on EHR data but the AUC dropped to a low results in incomplete datasets. | Missing data (e.g., smoking, BMI) impacts prediction accuracy. |
| Luo et al. (2024) [26] | Reviewed multimodal fusion techniques combining MRI, DTI, and EEG data to enhance stroke prediction. | Challenges in data alignment, registration, and high computational complexity. |
| Asiri et al. (2024) [16] | The Swin Transformer achieved high accuracy in tumor classification, applied to stroke detection as well. | High computational cost and small sample size limit clinical applicability. |
| Issaiy et al. (2025) [14] | Applied CNNs to predict hemorrhagic transformation after ischemic stroke. | Relies on large labeled datasets; limited generalizability due to dataset biases. |
| Zhang et al. (2025) [18] | Integrated 3D CNN with radiomics to differentiate stroke lesions from gliomas in MRI. Demonstrated highest specificity. | Reduced sensitivity in hyperacute strokes; relies on multi-parametric imaging. |
| Ou et al. (2025) [23] | Introduced a Transformer model using speech and movement data to identify strokes early. | Small dataset size and computational complexity limit generalizability. |
| Martín et al. (2025) [24] | Achieved the best accuracy in tumor classification using the Swin Transformer, applied to stroke detection as well. | High computational cost and limited generalizability due to small sample sizes. |



(a) Axial non-contrast CT scan (b) Axial CT brain from skull base to vertex

(c) Midline Coronal CT scan (d) Different Coronal CT brain anatomical views

**Fig. 2.** Sample dataset illustrating different CT imaging planes.

| Demographic details | | Biochemical markers | | | | | | Hematological parameters | | | Liver Enzymes | | Coagulation | Blood glucose levels | | Blood | Smoke | Ischamic |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Age | Gender | Cholestrol | HDL | LDL | Triglycerides | Creatinine | Urea | Hemoglobin | Platelete | WBC | ALT | AST | INR | FBS | Diabetes | Pressure | | Stroke |
| 27 | M | 150 | 55 | 90 | 100 | 0.78 | 9 | 13.9 | 240000 | 6500 | 11 | 18 | 1 | 70 | 0 | 120/80 | 0 | 0 |
| 64 | M | 230 | 40 | 130 | 160 | 1.3 | 31 | 13 | 220,000 | 7000 | 13 | 18 | 1.1 | 110 | 1 | 150/95 | 1 | 1 |
| 65 | M | 210 | 35 | 130 | 160 | 1.3 | 22 | 13.2 | 380,000 | 9500 | 55 | 45 | 1 | 120 | 1 | 150/90 | 1 | 1 |
| 74 | M | 240 | 38 | 150 | 188 | 1.45 | 36 | 11.9 | 290,000 | 8700 | 23 | 25 | 1.1 | 130 | 1 | 160/100 | 1 | 1 |
| 80 | M | 210 | 44 | 130 | 150 | 1.33 | 31 | 12.8 | 255,000 | 6000 | 22 | 22 | 1 | 139 | 1 | 150/95 | 1 | 0 |

**Fig. 3.** Sample patient medical record from the dataset.

**Table 2**
Overview of representative stroke-related datasets.

| Dataset (Reference) | # Records | EHR | CT |
|---|---|---|---|
| ISLES 2024 [38] | 250 | ✓ | ✓ |
| APIS [39] | 96 | ✓ | ✓ |
| CPAISD [40] | – | – | ✓ |
| PHE-SICH-CT-IDS [41] | 120 | – | ✓ |
| Multicenter AIS MRI [42] | 5788 | ✓ | – |
| MIMIC-III [43] | 53,423 | ✓ | – |
| ChestX-ray14 [44] | 112,120 | – | ✓ |
| CheXpert [45] | 224,316 | – | ✓ |
| MIMIC-CXR [46] | 377,110 | – | ✓ |
| OpenI [47] | 7470 | – | ✓ |
| PadChest [48,49] | 160,000+ | – | ✓ |
| COVID-19 Image Data Collection [50] | 950+ | – | ✓ |
| **HeteroStroke (Proposed)** | **530** | ✓ | ✓ |

## 4. Methodology

The proposed framework for stroke prediction is built upon multimodal data combining brain CT images and Electronic Health Records (EHR). The methodology comprises four core stages: data collection and preprocessing, feature extraction, multimodal fusion, and final classification. Three models were developed and compared: a baseline deep learning model on medical images Fig. 4, a multimodal fusion of medical images and clinical embeddings (HeteroFuseNet) Fig. 5, and meta-Learning with image models and clinical embeddings Fig. 6.

The proposed framework is succinctly summarized in Table 3. in a sequential manner, outlining the complete multimodal pipeline. The procedure initiates with the preparation of CT brain images and structured electronic health record (EHR) data. The CT pictures are normalized and resized to guarantee uniform input representation for the visual encoders. The clinical variables obtained from the EHR dataset undergo preprocessing, which includes normalization, encoding of categorical features, and addressing missing values.

After preprocessing, deep learning architectures are utilized to extract advanced imaging features from the CT images, while structured clinical features are processed via a specialized clinical branch. The retrieved representations are subsequently combined utilizing two complementary techniques. Initially, HeteroFuseNet executes feature-level fusion to integrate imaging and clinical representations into a cohesive feature space. Secondly, HeteroMetaNet employs a meta-learning-based decision fusion method that consolidates predictions produced by modality-specific models through a meta-classifier.

The consolidated predictions are employed to produce the definitive classification output for ischemic stroke forecasting. This pipeline enables the framework to efficiently acquire complementary information from both radiological and clinical data sources.

The HeteroMetaNet framework employs a meta-learning-inspired decision-level fusion technique tailored for the integration of heterogeneous multimodal inputs. In contrast to conventional ensemble or stacking methods that merge predictions from models utilizing analogous

**Table 3**
Overview of the proposed multimodal pipeline for ischemic stroke prediction.

| Pipeline stage | Description | Output |
|---|---|---|
| Data acquisition | Collection of CT brain images and structured electronic health record (EHR) variables from hospital clinical records. | Raw CT images and clinical data |
| Data preprocessing | CT images are normalized and resized, while clinical variables undergo normalization, categorical encoding, and missing value handling. | Cleaned CT images and structured clinical features |
| Feature extraction | Deep learning encoders (CNN and Transformer-based models) extract high-level imaging features from CT scans, while structured clinical features are processed through a dedicated clinical branch. | Imaging feature vectors and clinical feature representations |
| Feature-Level fusion (HeteroFuseNet) | Multimodal feature representations from CT imaging and EHR data are integrated into a unified feature space to capture complementary information across modalities. | Fused multimodal feature representation |
| Decision-Level fusion (HeteroMetaNet) | Predictions generated by modality-specific models are combined using a meta-learning strategy with a Random Forest meta-learner. | Aggregated prediction scores |
| Final prediction | The fused output is used to produce the final classification for ischemic stroke prediction. | Stroke prediction outcome |

feature representations, HeteroMetaNet consolidates predictions derived from diverse data modalities, such as CT imaging and structured electronic health record (EHR) data. Each modality-specific model encapsulates distinct facets of the clinical information domain. The meta-learner is subsequently trained to discern optimal weighting patterns among these diverse predictors, facilitating adaptive aggregation of modality-specific outputs. This design enables the framework to use cross-modal interactions that single-modality models or traditional ensemble methods may not completely capture.

The dataset was meticulously partitioned at the patient level before model training to guarantee reliable evaluation and avert potential data leakage. As each patient may have numerous CT slices, conducting the data split at the slice level could inadvertently result in slices from the same patient being included in both the training and testing sets. To mitigate this issue, all CT slices pertaining to a specific patient were exclusively allocated to either the training or testing group. This patient-specific partitioning technique guarantees that the evaluation process accurately simulates a true clinical deployment scenario, hence offering a more rigorous assessment of model generalization performance.

Furthermore, Principal Component Analysis (PCA) was utilized as a dimensionality reduction method to mitigate the high dimensionality and possible redundancy inherent in multimodal feature representations. PCA executes an effective linear transformation that retains the most significant variation in the data while considerably diminishing the number of features. In contrast to trainable dimensionality reduction methods like autoencoders, PCA does not incorporate extra trainable parameters, therefore preserving computing efficiency and minimizing the likelihood of overfitting in light of the moderate dataset size. This design decision fosters the creation of a lightweight and computationally efficient framework appropriate for real clinical settings.

### *4.1. Data collection and preprocessing*

The dataset, referred to as **HeteroStroke**, is a real-world collection sourced from Al Waha Hospital, comprising 530 anonymized patient records. Each record includes two key components: a structured electronic health record (EHR) file in a tabular format, containing demographic, clinical, and laboratory information, along with a folder containing 4 non-contrast CT brain images for each record.

All direct identifiers were removed to ensure patient confidentiality. Each patient was assigned a unique code used to associate their EHR data with their corresponding CT image folder. CT images were stored using standardized naming conventions for consistency, and EHR files were indexed using the same patient codes for accurate alignment during multimodal fusion.

To guarantee uniform data representation across modalities, many preprocessing processes were included for both the CT imaging data and the structured EHR information. The CT brain pictures were initially standardized by intensity normalization and scaled to a consistent spatial resolution appropriate for the input specifications of the deep learning architectures. Fundamental preprocessing techniques, such as noise reduction and image scaling, were implemented to enhance the stability of feature extraction.

Numerical variables in the structured clinical data were adjusted by z-score normalization to achieve equivalent feature scales, whereas categorical features were converted using one-hot encoding. Missing values in the EHR dataset were addressed by median imputation for numerical variables, ensuring resilience against outliers frequently present in clinical data.

During model training, deep learning models were optimized using typical stochastic gradient optimization with suitable learning rate adjustments and batch setups. The training procedure incorporated regularization techniques to avert overfitting and guarantee stable convergence. The preprocessing and training protocols were established to enhance the robustness and reproducibility of the multimodal framework.

The dataset is labeled into two classes: **Stroke** and **Non-Stroke**, with approximately 60% of the cases labeled as Stroke and 40% as Non-Stroke. This near-balanced distribution supports robust and unbiased training and evaluation of classification models.

***Ethical approval and informed consent***. All procedures involving human data in this study were reviewed and approved by the **Al Waha Hospital Institutional Review Board (IRB), Giza, Egypt**, on **26 August 2023**. The dataset provided by the hospital consisted of fully anonymized retrospective clinical and radiological records. In accordance with the IRB decision and applicable institutional guidelines, the requirement for informed consent was waived because no identifiable patient information was accessed, and the data were originally collected as part of routine clinical care. All methods were carried out in accordance with relevant guidelines and regulations, including the Declaration of Helsinki.

### *4.2. Feature extraction*

To enable cross-modal learning, high-level features were independently extracted from CT images and structured EHR data. These features were then integrated through two distinct fusion strategies: an image-only multi-backbone CT feature extraction fusion, combining outputs from ResNet-50, EfficientNetB0, and Swin Transformer architectures, and a comprehensive multimodal fusion that merges visual and clinical embeddings to enhance stroke prediction performance.

The CT imaging sector utilizes a multi-backbone feature extraction approach, employing three distinct deep learning architectures to derive varied visual representations from the identical CT modalities. Convolutional and transformer-based encoders are utilized concurrently on the CT slices to extract distinct spatial and contextual features of the brain images.

It is important to clarify that these branches do not pertain to distinct data modalities; instead, they signify various feature extraction backbones functioning on the identical CT imaging modality. The resultant feature representations are then combined to provide a more comprehensive and distinctive imaging representation prior to integration with the structured EHR characteristics.

### *4.3. CT image feature extraction*

To optimize feature extraction from CT images, both a uni-modal and a Multi-Backbone backbone architecture were employed. Rather than depending on a single model, the combination of three distinct architectures significantly improves generalizability and mitigates the limitations associated with individual models. This design choice is motivated by previous research demonstrating that integrating imaging features with clinical records substantially enhances predictive performance [51,52].

Each patient's CT images were processed using three pretrained models: two convolutional-based models (ResNet-50 and EfficientNetB0) and one transformer-based model (Swin Transformer). The models were utilized in standalone and tri-model architectures. All images were resized to 224 × 224 px to match the input requirements of the respective models.

*(1) ResNet-based feature extraction*

The first model employed is ResNet-50, a deep residual network with 50 layers utilizing skip connections to facilitate gradient flow and address vanishing gradient issues. ResNet-50 captures hierarchical visual features effectively, making it well-suited for analyzing medical images. The feature extraction process using ResNet50 can be mathematically expressed as Eq. (1):

$$F_{\text{ResNet}} = \text{ReLU}(W_l \cdot F_{l-1} + b_l) \tag{1}$$

where $W_l$ and $b_l$ are the weights and biases of the layer $l = 50$ layers, respectively, and $F_{l-1}$ is the feature map from the previous layer.

*(2) EfficientNet-based feature extraction*

The second model is EfficientNetB0, known for achieving high representational capacity while maintaining computational efficiency. It employs a compound scaling method that uniformly scales network depth, width, and input resolution. The output feature map $F_{\text{EfficientNet}}$ from EfficientNetB0 for an input image $X$ is given by Eq. (2):

$$F_{\text{EfficientNet}} = f_{\theta}^{\text{EffNet}}(X) \tag{2}$$

where $f_{\theta}^{\text{EffNet}}$ denotes the EfficientNetB0 function parameterized by weights $\theta$. Extracted features are then passed through a global average pooling layer and flattened to produce a compact vector representation suitable for multimodal fusion.

*(3) transformer-based feature extraction*

The third model used is the Swin Transformer, a vision transformer that processes images using shifted windows, capturing both local and global contextual information efficiently. Unlike convolutional operations in CNNs, the Swin Transformer relies on self-attention mechanisms defined as Eq. (3):

$$F_{\text{Swin}} = \text{sigmoid}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{3}$$

where $Q$, $K$, and $V$ represent the query, key, and value matrices, respectively, and $d_k$ is the scaling factor. The Swin Transformer's capacity to model long-range dependencies makes it particularly effective for complex medical imaging tasks.

Feature vectors were extracted from the final convolutional or transformer layers of each model using global average pooling. Given that multiple CT slices were available per patient, mean pooling was applied across slices to yield a single representative feature vector per model for each patient.

In addition to evaluating each model individually, a Multi-Backbone fusion strategy was implemented to combine image-only features. As illustrated in Fig. 4, feature vectors extracted from ResNet-50, EfficientNetB0, and Swin Transformer were concatenated to form a unified visual embedding. Prior to concatenation, Principal Component Analysis (PCA) was applied exclusively to the extracted image features, retaining 95% of the original variance. This dimensionality reduction step automatically determined the optimal number of components required for efficient image-only fusion while minimizing information loss. Consistent with previous studies, the application of PCA enhanced computational efficiency, reduced feature redundancy, and improved the quality of feature integration in medical imaging tasks [28–30]. The resulting fused image representation was subsequently passed through a fully connected neural network comprising two hidden layers with 128 and 64 neurons, respectively, to generate the final stroke prediction. Additionally, the distribution of the latent embeddings confirmed that the fused features enabled clear class separation, thereby demonstrating the effectiveness of the image-only fusion strategy.

This dual-phase setup enabled a comparative assessment between the performance of stand-alone models and the fused image-only strategy, providing valuable insights into the advantages of model-level integration within a unimodal context.

#### *4.3.1. EHR feature extraction*

The structured electronic health record (EHR) data were preprocessed prior to integration with the imaging features. Numerical attributes, including age, systolic blood pressure, and fasting blood glucose, were standardized using z-score normalization to ensure consistent feature scaling. Categorical variables, if present, were transformed using one-hot encoding.

The standardized EHR feature vector was subsequently passed through a multilayer perceptron (MLP) comprising two hidden layers with 128 and 64 neurons, respectively, each followed by ReLU activation functions. The output of the second layer was then projected into a 128-dimensional dense representation, forming the clinical embedding $F_{\text{clinical}}$. This transformation process is expressed mathematically in Eq. (4).

$$F_{\text{clinical}} = \text{ReLU}\left(W_2\, \text{ReLU}\left(W_1\, \hat{X}_{\text{EHR}} + b_1\right) + b_2\right) \tag{4}$$

In Eq. (4), $\hat{X}_{\text{EHR}}$ denotes the standardized EHR feature vector containing key clinical indicators such as age, systolic blood pressure, and fasting blood glucose. $W_1$ and $W_2$ represent the weight matrices of the first and second hidden layers, while $b_1$ and $b_2$ denote their corresponding bias terms. The ReLU activation introduces non-linearity, enabling the model to capture complex physiological relationships within the data. The resulting embedding $F_{\text{clinical}} \in \mathbb{R}^{128}$ encapsulates a compact and discriminative clinical representation for each patient.

This EHR encoding process is visualized in both the model-level fusion (*HeteroFuseNet*) and meta-learning (*HeteroMetaNet*) architectures illustrated in Figs. 5 and 6, respectively.

### *4.4. Multimodal fusion strategies*

Two multimodal fusion strategies were implemented. The first, *HeteroFuseNet*, integrates heterogeneous data sources — CT brain images and structured EHR — within a unified deep learning architecture. This model performs model-level fusion by concatenating image and clinical feature vectors at the feature level, followed by joint representation learning through a multilayer perceptron. The architecture is designed to capture the complementary nature of visual and clinical information to enhance stroke prediction performance.

The second strategy, HeteroMetaNet, applies a meta-learning approach for late-level fusion. Instead of mid-level feature integration,

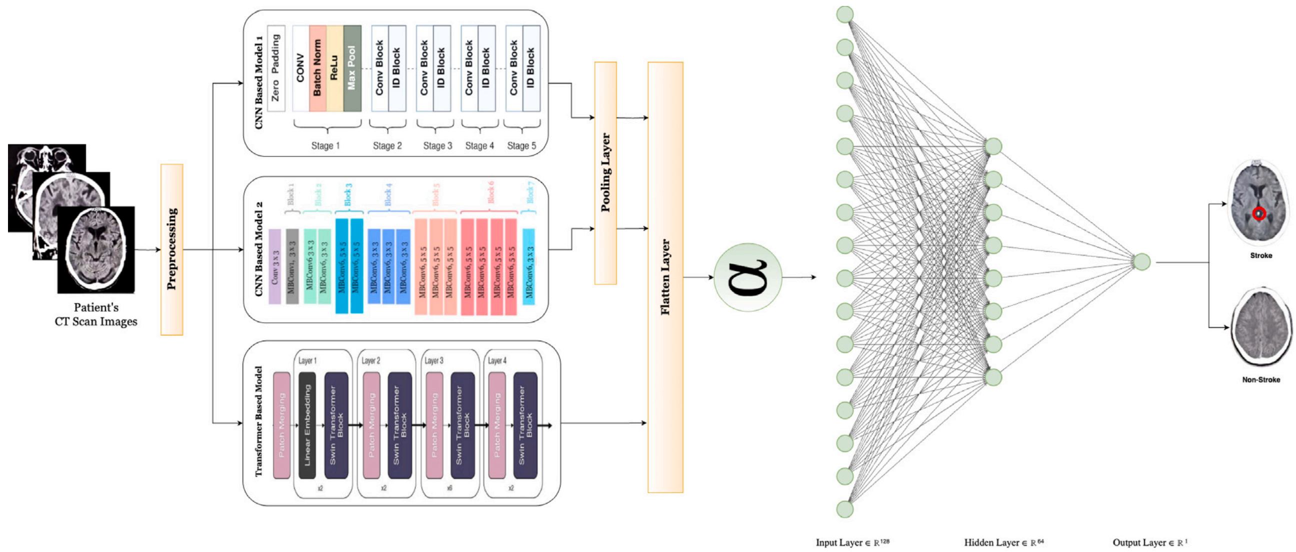



**Fig. 4.** Architecture for single-modality evaluation.

this model employs late fusion by combining the outputs of the individual image-based learners (ResNet-50, EfficientNetB0, and Swin Transformer) along with the structured EHR embedding. Prior to training the individual image models, Principal Component Analysis (PCA) was applied to the extracted image features to enhance class separability and reduce dimensionality. A meta-classifier—specifically, a Random Forest model with 100 estimators—is then used to process the concatenated outputs and produce the final stroke prediction. This hierarchical fusion approach aims to improve generalization by adaptively weighting the contributions of each modality, leveraging both imaging and clinical perspectives.

The Random Forest meta-learner was selected due to its robustness to overfitting, ability to handle heterogeneous feature distributions, and strong performance in ensemble learning settings. Compared to alternative meta-classifiers such as logistic regression or gradient boosting, Random Forest provides a balance between interpretability, computational efficiency, and non-linear decision modeling, making it particularly suitable for aggregating multimodal predictions.

### 4.4.1. HeteroFuseNet: Model-level fusion

HeteroFuseNet, illustrated in Fig. 5, performs mid-level multimodal fusion between visual and clinical features for stroke prediction. Image features are independently extracted from three vision encoders—ResNet-50, EfficientNetB0, and Swin Transformer—as formulated in Eqs. (1)–(3). To reduce redundancy and enhance computational efficiency, Principal Component Analysis (PCA) is applied to each encoder's output before fusion. The resulting reduced-dimensional feature sets are flattened and concatenated to form a unified visual representation, as shown in Eq. (5):

$$F_{\text{image}} = \text{Flatten}\left( [\,\text{PCA}(F_{\text{ResNet}}), \text{PCA}(F_{\text{EfficientNet}}), \text{PCA}(F_{\text{Swin}})\,] \right) \tag{5}$$

The EHR branch processes structured variables such as age, systolic blood pressure, and fasting blood glucose. These inputs are standardized using z-score normalization and encoded through a two-layer neural network with 128 and 64 neurons, respectively, to produce a compact clinical embedding $F_{\text{clinical}} \in \mathbb{R}^{128}$ that encapsulates patient-specific physiological information.

The image (5) and clinical embeddings (4) are then concatenated to form the multimodal feature vector $\boldsymbol{\alpha}$, integrating both modalities as expressed in Eq. (6):

$$\boldsymbol{\alpha} = \text{Concat}(F_{\text{image}}, F_{\text{clinical}}) \tag{6}$$

Finally, $\boldsymbol{\alpha}$ is passed through a two-layer classifier followed by a sigmoid activation to produce the ischemic stroke prediction probability $\hat{y}$, as given in Eq. (7):

$$\hat{y} = \sigma\big(\text{NN}_{\text{classifier}}(\boldsymbol{\alpha})\big) \tag{7}$$

Here, $\sigma$ denotes the sigmoid function and $\hat{y} \in [0,1]$ represents the predicted likelihood of ischemic stroke. The fusion of visual and clinical embeddings enhances the model's discriminative capability, as confirmed by latent feature visualization showing improved class separability.

### 4.4.2. HeteroMetaNet: Late-level fusion

The HeteroMetaNet model, illustrated in Fig. 6, employs a late-level fusion strategy following a meta-learning paradigm. By employing a late fusion strategy, HeteroMetaNet integrates modality-specific predictions rather than raw features, allowing the meta-learner to capture complementary decision patterns across imaging and clinical modalities. This design enhances generalization and robustness, particularly when modality-specific noise or feature imbalance is present. This architecture captures higher-order dependencies by aggregating the predictions of multiple weak learners. Each of the three image-based feature vectors, extracted respectively from ResNet-50, EfficientNetB0, and Swin Transformer, is passed through a dedicated weak learner—a fully connected neural network comprising two hidden layers (with 128 and 64 neurons, respectively) and a final sigmoid output neuron. Dropout layers (with a rate of 0.3) follow each hidden layer for regularization and overfitting mitigation.

The outputs from these three weak learners, $\hat{y}_{\text{ResNet}}$, $\hat{y}_{\text{EffNet}}$, and $\hat{y}_{\text{Swin}}$, represent the modality-specific predictions derived from imaging data. Before meta-level fusion, Principal Component Analysis (PCA) is applied to the set of image-driven outputs.

Simultaneously, the EHR data is embedded using the same structured processing pipeline described in Section 4.4.1, producing an EHR feature vector $F_{\text{clinical}} \in \mathbb{R}^{128}$.

The final meta-representation $F_{\text{meta}}$ is formed by concatenating the PCA-transformed image outputs with the clinical embedding, as expressed in Eq. (8):

$$F_{\text{meta}} = [\text{PCA}([\hat{y}_{\text{ResNet}}, \hat{y}_{\text{EffNet}}, \hat{y}_{\text{Swin}}]), F_{\text{clinical}}] \tag{8}$$

This concatenated feature vector $F_{\text{meta}}$ is subsequently passed to a final classification network—referred to as the meta-learner—implemented using a Random Forest (RF) model with 100 estimators.

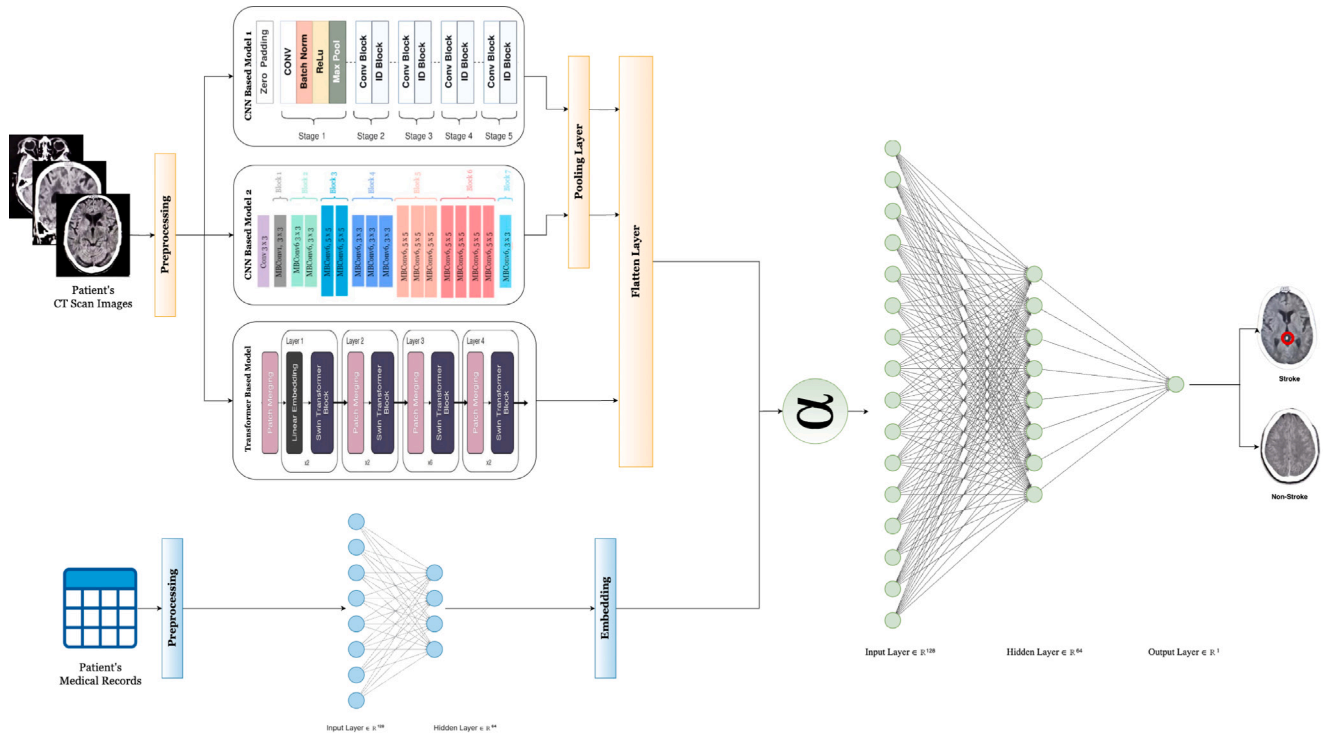



**Fig. 5.** Architecture of HeteroFuseNet for multimodal stroke prediction.

**Table 4**
Comparison of HeteroMetaNet with existing meta-learning frameworks.

| Aspect | MAML | MMAML | MLDG | HeteroMetaNet |
|---|---|---|---|---|
| Meta-learning objective | Few-shot adaptation | Task + modality adaptation | Domain generalization | Decision-level fusion |
| Training paradigm | Episodic | Episodic + modality tokens | Meta-domain split | Supervised + stacking |
| Modality handling | Single | Multi (late) | Single | Multi (late fusion) |
| Dimensionality reduction | No | No | No | PCA pre-fusion |
| Meta-learner | Gradient (SGD) | Gradient (SGD) | Gradient (SGD) | Random Forest |
| Backprop through meta-learner | Yes | Yes | Yes | No |
| Interpretability | Low | Low | Low | High |

The meta-learner is trained to optimally integrate modality-specific outputs and clinical data to generate the final stroke prediction $\hat{y}_{\text{final}}$, as defined in Eq. (9):

$$\hat{y}_{\text{final}} = \text{MetaLearner}(F_{\text{meta}}) \tag{9}$$

HeteroMetaNet differs fundamentally from gradient-based meta-learning frameworks such as Model-Agnostic Meta-Learning (MAML), Multimodal MAML (MMAML), and Meta-Learning for Domain Generalization (MLDG). Table 4 summarizes these distinctions.

Importantly, HeteroMetaNet is not a gradient-based meta-learning method but rather a stacking ensemble with a meta-learner—a conceptual distinction we explicitly emphasize. While MAML and its variants focus on rapid adaptation to new tasks via episodic training and gradient-based optimization, HeteroMetaNet addresses a different problem: the stable integration of heterogeneous modality-specific predictions from fixed, pre-trained encoders. Our approach prioritizes robustness, interpretability, and computational efficiency over few-shot adaptability, making it more suitable for clinical deployment where training data are moderately sized but heterogeneous.

The HeteroMetaNet training procedure comprises two distinct stages with no gradient flow between them.

*Stage 1: Modality-specific training.* Each image backbone (ResNet-50, EfficientNetB0, Swin Transformer) is paired with an identical weak learner architecture—a fully connected network with two hidden layers (128 and 64 units, ReLU activations, dropout 0.3) and a sigmoid output. Each weak learner is trained independently using binary cross-entropy loss as defined in Eq. (10):

$$\mathcal{L}_i = -\frac{1}{N}\sum_{j=1}^{N}\left[y_j \log(\hat{y}_{i,j}) + (1-y_j)\log(1-\hat{y}_{i,j})\right] \tag{10}$$

where $i \in \{\text{ResNet}, \text{EfficientNet}, \text{Swin}\}$ indexes the backbone, $j$ indexes the patient, $y_j$ is the ground-truth label, and $\hat{y}_{i,j}$ is the predicted probability. No weight sharing or cross-modal information is exchanged during this stage. The structured EHR branch is trained simultaneously using the same binary cross-entropy loss, producing the clinical embedding $F_{\text{clinical}}$ and an auxiliary prediction $\hat{y}_{\text{EHR}}$ (though the EHR prediction is not used in the final meta-representation; only the embedding $F_{\text{clinical}}$ is retained).

*Stage 2: Meta-learner training.* The meta-feature vector $F_{\text{meta}}$ is constructed in Eq. (11):

$$F_{\text{meta}} = \left[\text{PCA}\left(\left[\hat{y}_{\text{ResNet}}, \hat{y}_{\text{EfficientNet}}, \hat{y}_{\text{Swin}}\right]\right), F_{\text{clinical}}\right] \tag{11}$$

PCA is fitted on the training set's image predictions and applied to both training and test sets, retaining 95% variance (automatically determining the number of components, typically 2–3). The Random Forest meta-learner is then trained on $\{F_{\text{meta}}^{(j)}, y_j\}$ using standard entropy-based splitting criteria. Crucially, the meta-learner does not backpropagate gradients to the modality-specific models, which prevents any single modality from dominating the learning signal and allows each encoder to specialize in its respective data type.

Random Forest was chosen as the meta-learning classifier because of its robustness in modeling nonlinear interactions and its capacity to successfully incorporate diverse prediction variables produced by the modality-specific models. Moreover, Random Forest exhibits robust generalization capabilities and necessitates minimal hyperparameter optimization, rendering it particularly appropriate for the late fusion phase of the proposed multimodal framework. This hierarchical meta-learning design enables the model to benefit from both modality-specific confidence scores and patient-specific clinical context, supporting more accurate and generalizable stroke prediction across heterogeneous inputs.

Crucially, HeteroMetaNet must not be construed as a traditional gradient-based meta-learning framework like Model-Agnostic Meta-Learning (MAML). Conventional meta-learning methodologies are often formulated for fast adaptation across many tasks through episodic training and inner–outer optimization cycles. Conversely, the proposed system targets a distinct aim: resilient multimodal decision integration across diverse clinical and radiological data sources.

Instead of acquiring starting parameters for rapid adaptation, HeteroMetaNet functions as a meta-learning-inspired stacking architecture where modality-specific predictors autonomously learn complementary representations from CT imaging and structured electronic health record data. Their results are subsequently consolidated via a superior meta-learner to enhance classification robustness and generalization. Consequently, the framework emphasizes adaptive decision fusion, interpretability, and computational efficiency rather than few-shot task adaptation.

In contrast to traditional meta-learning approaches that emphasize swift adaptation across task distributions, the HeteroMetaNet framework centers on the decision-level integration of diverse modality-specific predictors. Each modality-specific branch autonomously acquires supplemental diagnostic data from CT imaging and structured electronic health record variables. The Random Forest meta-learner subsequently consolidates these predictive signals to provide a resilient final conclusion. This technique facilitates efficient multimodal integration while preserving computational efficacy and enhanced interpretability, which are crucial in practical clinical deployment contexts.

## 5. Evaluations and experiments

This section presents the experimental design, implementation details, and evaluation results used to validate the proposed multimodal stroke prediction framework. The experiments were structured to systematically assess the effectiveness of each model component and the contribution of multimodal integration.

First, single-modality experiments were conducted on CT brain images to establish baseline performance using three state-of-the-art visual encoders: ResNet-50, EfficientNetB0, and Swin Transformer. These models served as strong visual baselines for comparison.

Subsequently, multimodal experiments were performed to explore the synergistic effects of combining imaging features with structured clinical records (EHR). Two distinct architectures — HeteroFuseNet and HeteroMetaNet — were developed to evaluate feature-level and meta-level fusion strategies, respectively.

An ablation-style evaluation was performed across various configurations to substantiate the architectural design decisions, encompassing single-modality image models, image-only tri-backbone fusion, multimodal feature-level fusion (HeteroFuseNet), and meta-learning-based late fusion (HeteroMetaNet). This iterative assessment facilitates the examination of each component's contribution within the proposed framework, illustrating the performance enhancements realized via multimodal integration and meta-learning. The section also details the computational setup, hyperparameter configuration, evaluation metrics, and data partitioning protocols used to ensure reproducibility and fairness across all experimental settings. Comparative analyses are then presented to demonstrate how multimodal learning improves prediction accuracy, interpretability, and generalization in ischemic stroke detection.

To strengthen the reliability of the evaluation results, confidence intervals were calculated for the main performance metrics, including accuracy, F1-score, and ROC-AUC. These measurements offer an assessment of performance variability and reinforce the robustness of the proposed multimodal framework in comparison to unimodal baselines.

### 5.1. Experimental setup

All experiments reported in this study were conducted locally using a workstation equipped with dual Intel® Xeon® CPU E5-2683 v3 processors operating at 2.00 GHz and 96 GB RAM under Windows 11 Pro (64-bit). Model implementation and training were performed using Python 3.11 with TensorFlow and Keras. The same computational environment was consistently used across all single-modality and multimodal experiments, including HeteroFuseNet and HeteroMetaNet, to ensure fair comparison and reproducibility of the reported results. The proposed framework was designed to balance predictive performance and computational efficiency while remaining practical for deployment in clinical research environments. The reported experiments demonstrate that the framework can be executed using a standard workstation configuration, supporting reproducibility and accessibility without reliance on large-scale computational resources.

### 5.2. Hyperparameters configuration

All models were trained for 100 epochs using a batch size of 4. The Adam optimizer was adopted for its adaptive learning capabilities, with the learning rate set to the default value of 0.001. The binary cross-entropy loss function was employed to optimize the binary classification task, and a sigmoid activation function was used at the output layer to yield probability scores for stroke prediction. Input CT images were resized to ($224 \times 224$) px, depending on the model, and normalized to a range of [0, 1] before being passed to the network. Early stopping was utilized to prevent overfitting, monitoring the validation loss with a patience threshold of 10 epochs. The best-performing model weights during training were preserved using a model checkpointing strategy.

Principal Component Analysis (PCA) was utilized to diminish feature dimensionality and alleviate redundancy across diverse feature representations. PCA enhances computational efficiency and facilitates stable multimodal integration by retaining the most informative elements and eliminating noise and correlations.

For the model-level fusion model (*HeteroFuseNet*), image features were independently extracted from ResNet-50, EfficientNetB0, and Swin Transformer. PCA was applied separately to each set of extracted features for dimensionality reduction while retaining 95% of the variance. The PCA-reduced features were flattened and concatenated into a unified visual embedding. Structured EHR features, processed through a two-layer neural subnetwork with 128 and 64 neurons using ReLU activations, were then concatenated with the visual embedding. The combined multimodal representation was passed through an additional two-layer MLP before the final sigmoid classification layer to predict stroke risk.

For the meta-learning model (*HeteroMetaNet*), the outputs of the image-based weak learners were concatenated and subjected to a single PCA transformation. The reduced representation was then combined with the structured EHR embedding and passed to a Random Forest meta-learner with 100 estimators to generate the final stroke prediction.

The main training configurations and preprocessing strategies applied across all experiments, including both single-modality and multimodal settings, are summarized in Table 5. These standardized settings were implemented to ensure consistency and fairness, allowing for a reliable comparison of performance across different architectures.

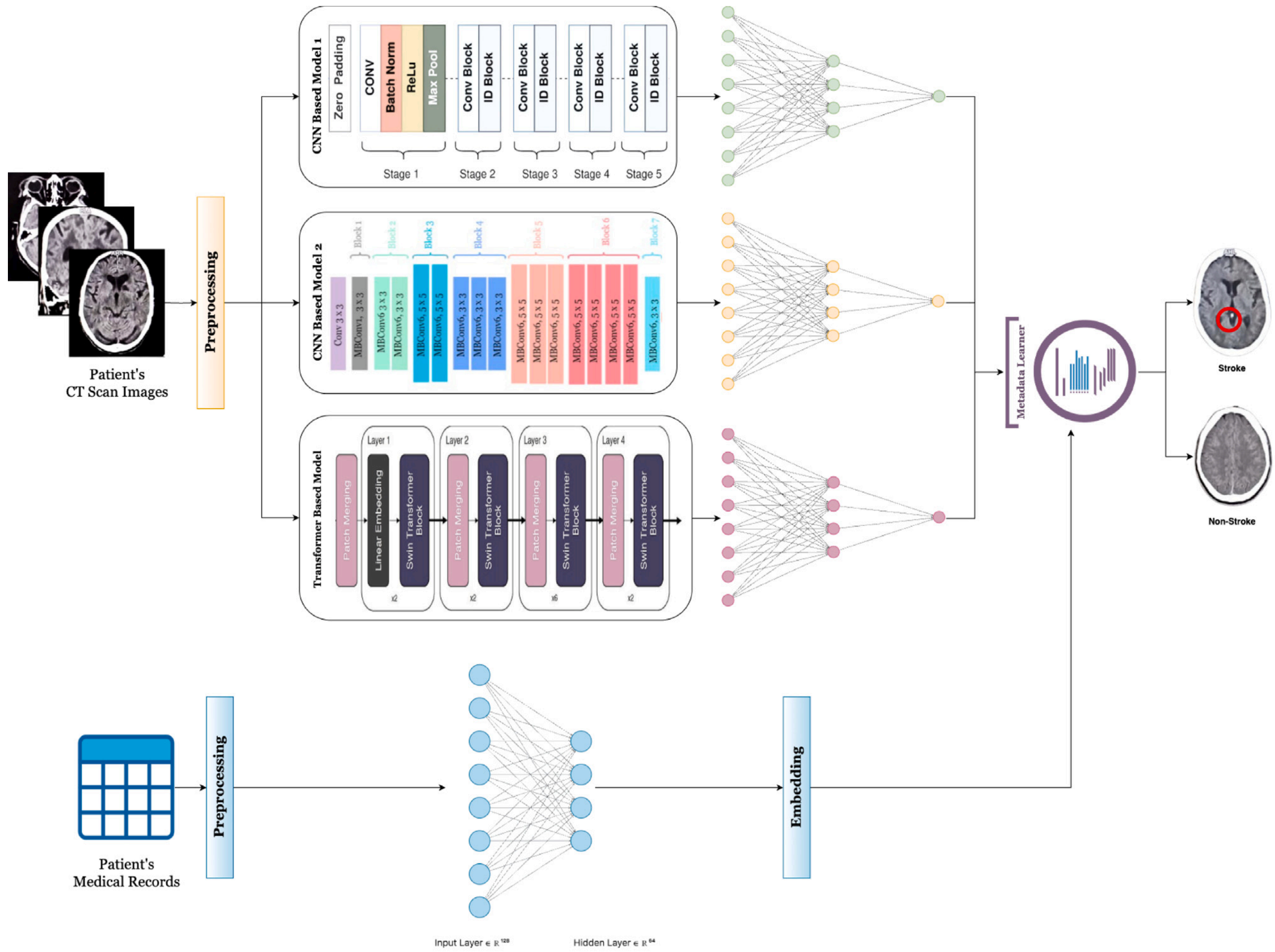



**Fig. 6.** Architecture of HeteroMetaNet for multimodal stroke prediction.

**Table 5**
Constant parameters across all experiments.

| Parameter | Value |
|---|---|
| Activation | ReLU (Rectified Linear Unit) |
| Batch size | 4 |
| Epochs | 100 with early stopping (patience = 10) |
| Loss | Binary Cross-Entropy |
| Optimizer | Adam |
| Learning rate | 0.001 |
| Dropout rate | 0.3 |
| Dense layers | 2 |
| Dense units | 128, 64 |
| Output layer | Sigmoid |

### *5.3. Computational complexity and reproducibility*

The computational complexity of the suggested framework depends on the utilized backbone architecture. ResNet-50 and EfficientNetB0 predominantly utilize convolutional operations, with computational complexity dependent on spatial resolution and convolution kernel dimensions, whereas the Swin Transformer employs window-based self-attention mechanisms that exhibit diminished quadratic complexity relative to traditional Vision Transformers. To alleviate computational burden and feature redundancy, Principal Component Analysis (PCA) was employed before fusion, significantly decreasing the dimensionality of multimodal representations while retaining 95% of the original variance.

The HeteroMetaNet design enhances performance by utilizing a lightweight Random Forest meta-learner that functions on low-dimensional prediction vectors instead of raw feature maps. This architecture markedly decreases memory usage and training expenses relative to end-to-end multimodal transformers. All computational complexity analysis and reproducibility experiments were conducted using the same hardware and software environment described in Section 5.1 to maintain consistency throughout the study. The experiments were executed on a workstation equipped with dual Intel® Xeon® CPU E5-2683 v3 processors operating at 2.00 GHz and 96 GB RAM under Windows 11 Pro (64-bit).

This unified environment ensured reliable benchmarking across all architectures and eliminated ambiguity regarding the distinction between model training, evaluation, and computational analysis.

To ensure reproducibility and openness, the implementation code, preprocessing scripts, and experimental setup will be made publicly accessible upon the manuscript's approval.

### *5.4. Evaluation metrics*

To ensure the robustness of our comparative evaluations, we computed 95% confidence intervals (CIs) for all primary performance metrics (accuracy, precision, recall, F1-score, and ROC-AUC) using bootstrap resampling with 1000 iterations. For each iteration, the test set was resampled with replacement to the original test set size, and metrics were recalculated. The 2.5th and 97.5th percentiles of the bootstrap distribution defined the CI bounds. Additionally, pairwise statistical significance between models was assessed using McNemar's test for matched proportions, with p-values $< 0.05$ considered statistically significant. These analyses provide a quantitative assessment of performance variability and reinforce the robustness of observed improvements over unimodal baselines.

Model performance was assessed using several metrics tailored for binary classification tasks. The dataset was divided into 80% for training and 20% for testing. Accuracy was employed to measure the overall proportion of correctly predicted instances. Precision reflected the model's ability to minimize false positives, while recall captured its sensitivity to actual positive cases. The F1-score, calculated as the harmonic mean of precision and recall, offered a balanced evaluation, particularly valuable in scenarios where class distributions may vary. Additionally, the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) was used to assess the model's capacity to discriminate between the two classes—Stroke and Non-Stroke—across various threshold settings. To avert data leaking between the training and assessment phases, dataset partitioning was executed at the patient level instead of the image level. Each patient folder comprised several CT slices, with all slices associated with a particular patient allocated solely to either the training or testing set. As a result, no photos from an individual patient were present in both subsets. The dataset was partitioned into 80% for training and 20% for testing patients. Dataset partitioning was executed at the patient level to avert data leakage between the training and testing sets. Each patient folder comprises multiple CT slices, with all slices pertaining to a single patient exclusively allocated to either the training or testing subset. As a result, no patient provided photos for both subsets.

Missing values in the clinical dataset were addressed by median imputation for numerical variables, a method that is resilient to outliers and frequently employed in clinical machine learning applications. Furthermore, categorical clinical attributes were transformed by one-hot encoding, whilst numerical variables underwent z-score normalization to maintain uniform feature scaling prior to integration with picture features. Visual comparisons of model performance are presented in the corresponding results section along with references to the three model structures illustrated in Figs. 4–6.

### 5.5. Evaluation of single modality on medical images

In the first experimental setup, the predictive performance of CT brain images alone for ischemic stroke classification was assessed. All input images were first subjected to standardized preprocessing procedures, including intensity normalization and resizing, to ensure consistency and compatibility across models.

Initially, three deep learning architectures—ResNet-50, EfficientNetB0, and the Swin Transformer—were independently trained as strong single-modality baselines. These models were chosen for their proven ability to extract diverse hierarchical features, capturing both local and global visual clues essential for stroke identification. Preprocessing pipelines were adapted to match the specific input constraints of each model to optimize learning.

Besides evaluating the models independently, the tri-backbone evaluation explored involving image-only Multi-Backbone fusion, as shown in Fig. 4. This experiment mirrors the multimodal fusion strategy presented later in this study but focuses exclusively on visual data. This fused representation was passed through a fully connected network to generate the final stroke prediction. This fusion approach allows the model to use the best features of both the CNN-based and Transformer-based architectures within a single-modality framework.

This dual-phase setup allowed for a comparative assessment between stand-alone model performance and the fused image-only strategy, providing insights into the benefits of model-level integration within a unimodal context.

### 5.6. Evaluation of multimodality integration with CT imaging and EHR

To evaluate the added value of multimodal data in stroke prediction, two experimental setups were designed that jointly utilize brain CT images and structured electronic health records (EHR): **HeteroFuseNet** and **HeteroMetaNet**. These experiments aim to assess the complementary impact of combining visual and clinical data at both the feature and decision levels.

HeteroFuseNet, illustrated in Fig. 5, implements a model-level fusion strategy via a dual-branch architecture. In the imaging branch, CT scans are preprocessed and passed through three backbone deep learning encoders—ResNet-50, EfficientNetB0, and Swin Transformer. Each model captures diverse spatial representations of the brain, which are then flattened and concatenated to construct a comprehensive visual embedding.

In contrast, HeteroMetaNet, shown in Fig. 6, adopts a late-level fusion approach rooted in meta-learning. After generating intermediate predictions from each of the three baseline visual models via individual weak learners, these scalar outputs are combined with the EHR embedding into a compact vector denoted as $F_{\text{meta}}$ in Eq. (8). This composite representation is passed to a meta-learner—Random Forest classifier—to produce the final binary prediction.

This hierarchical configuration enables the model to learn modality-specific decision boundaries before integrating them contextually with clinical signals. As with HeteroFuseNet, PCA visualizations of the meta-representations demonstrate separability between stroke and non-stroke samples, validating the benefit of late fusion strategies.

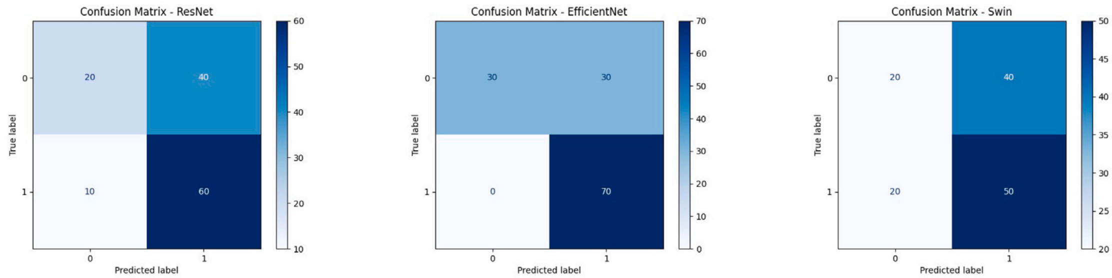
Overall, the results across these two experiments underscore the superior performance of multimodal architectures relative to single-modality baselines. The consistent gains observed with both model-level (HeteroFuseNet) and late-level (HeteroMetaNet) fusion confirm that EHR data enhances imaging characteristics by providing essential contextual information, ultimately enhancing the robustness and clinical applicability of stroke prediction models.

## 6. Results and findings

The authors assessed the computing efficiency of the proposed framework by examining the approximate model complexity and inference demands. The chosen backbone designs are comparatively lightweight in relation to extensive transformer-based multimodal systems. Furthermore, the PCA-based dimensionality reduction phase decreases feature dimensionality prior to fusion, The Random Forest meta-learner also provides an opportunity for interpretability through feature importance analysis. By examining the contribution of clinical variables and model predictions to the final decision, clinicians may gain insights into which factors most strongly influence stroke risk predictions. These design decisions facilitate a balanced model complexity while ensuring robust prediction performance, hence enhancing the framework's potential utility in resource-limited clinical settings.

The experimental design assesses the efficacy of the proposed multimodal architectures by comparing them with several unimodal baseline models that are solely trained on CT imaging features. These baselines serve as reference points for evaluating the efficacy of multimodal fusion techniques. The comparison between feature-level fusion (HeteroFuseNet) and decision-level fusion (HeteroMetaNet) facilitates a systematic analysis of the effects of various integration strategies within a unified experimental framework.

The studies involve a systematic assessment of three configurations to investigate the contribution of each component of the proposed framework: image-only models, multimodal feature-level fusion (HeteroFuseNet), and meta-learning-based late fusion (HeteroMetaNet). This study presents an ablation analysis that emphasizes the performance enhancements achieved via multimodal integration and meta-level fusion. Each experiment was assessed using accuracy, precision, recall, and F1-score as primary performance metrics.




(a) ResNet-50 (b) EfficientNetB0 (c) Swin Transformer

**Fig. 7.** Confusion matrices for baseline models: (a) ResNet-50, (b) EfficientNetB0, & (c) Swin Transformer.

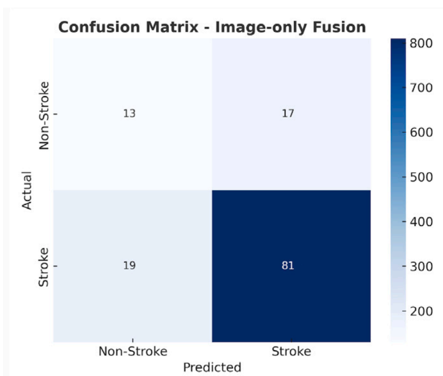



**Fig. 8.** Confusion matrix for image-only multi-backbone fusion model.

### 6.1. Evaluation 1 results: Single modality and image-only multi-backbone fusion

The first set of experiments focused on evaluating the individual performance of three deep learning models—ResNet-50, EfficientNetB0, and Swin Transformer—when trained exclusively on CT brain images. These models were used to establish baseline results using only visual features. ResNet-50 achieved an accuracy of 61.54%, a precision of 63.08%, a recall of 61.54%, and an F1-score of 58.52%. EfficientNetB0 outperformed the other vision models, yielding an accuracy of 76.92%, a precision of 83.85%, a recall of 76.92%, and an F1-score of 75.11%. In contrast, the Swin Transformer struggled to differentiate between stroke and non-stroke classes, achieving an accuracy of 53.85%, a precision of 52.99%, a recall of 53.85%, and an F1-score of 52.12%. The confusion matrices for each individual model are illustrated in Fig. 7. To enhance visual representations, a second experiment—referred to as Image-only Multi-Backbone fusion in evaluation Section 5.5—was conducted by concatenating the feature vectors extracted from the three individual models. This fusion approach led to noticeable performance improvements, achieving an accuracy of 80.77%, a precision of 82.35%, a recall of 80.77%, and an F1-score of 80.47%. The confusion matrix for the image-only multi-backbone fusion model is illustrated in Fig. 8, providing further insights into their classification patterns and misclassification tendencies.

While the fusion of visual features improved performance over individual models, relying solely on image data still presents critical limitations—particularly in complex clinical settings where contextual medical history is essential. These gaps emphasize the need to incorporate complementary structured data from EHRs, which motivates the multi-backbone evaluation described in the following subsection.

**Table 6**
Pairwise statistical significance using McNemar's test ($\alpha = 0.05$).

| Comparison | $p$-value | Significant? |
|---|---|---|
| HeteroMetaNet vs. EfficientNetB0 | 0.0021 | Yes |
| HeteroMetaNet vs. Image-only Fusion | 0.0412 | Yes |
| HeteroMetaNet vs. HeteroFuseNet | 0.0314 | Yes |
| HeteroFuseNet vs. EfficientNetB0 | 0.0873 | No (trend only) |

### 6.2. Evaluation 2 results: Multimodal fusion using clinical and imaging features

The second evaluation focused on assessing the performance of multimodal fusion strategies that combine extracted features from CT brain images with structured clinical records (EHR). This evaluation encompassed two complementary models: HeteroFuseNet, which employs model-level fusion, and HeteroMetaNet, which leverages late-level meta-learning.

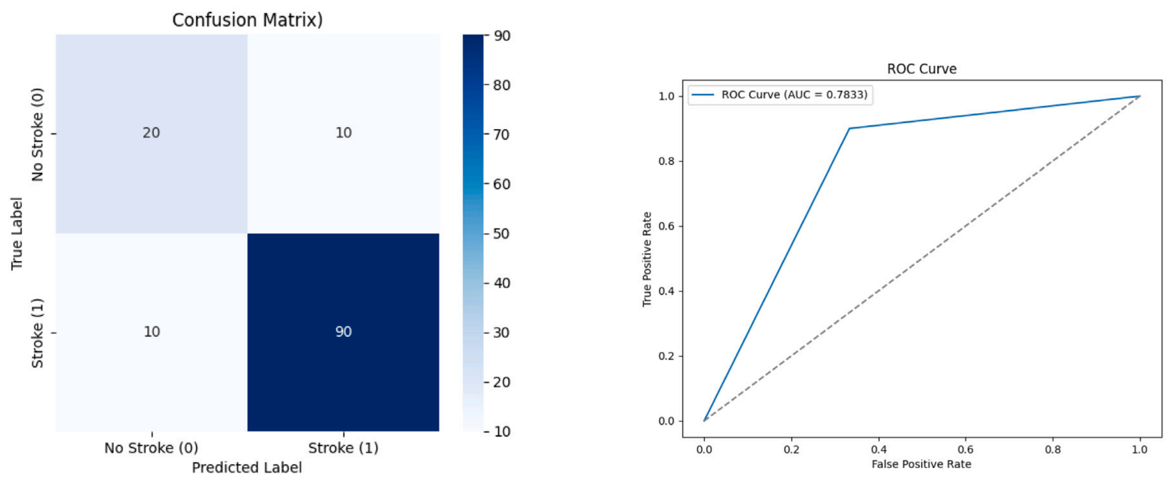
This experimental setup evaluated the performance of the proposed HeteroFuseNet model, which integrates CT-based visual embeddings with structured clinical data at the feature level. The model achieved robust classification performance, with an accuracy of 84.62%, precision of 90.00%, recall of 90.00%, and F1-score of 90.00%. However, the ROC-AUC was slightly lower at 78.33%, suggesting that the model, while accurate and consistent in prediction, may still have potential for improvement in threshold-based discrimination. The confusion matrix (a) and ROC curve (b) presented in Fig. 9 provide visual representations of the model's prediction quality and sensitivity-specificity tradeoff.

The final experiment applied meta-learning to combine the outputs of three image models alongside embeddings from structured clinical features. The HeteroMetaNet model achieved the highest performance across all configurations, with an accuracy of 92.31%, precision of 100%, recall of 90.91%, F1-score of 95.24%, and ROC-AUC of 95.45%. The high precision and ROC-AUC values emphasize the model's ability to minimize false positives and effectively separate stroke from non-stroke cases. Confusion matrix and ROC curve visualizations for this model are shown as (a) and (b), respectively, in Fig. 10.

This comparative analysis highlights the significant advantages of meta-learning architectures in employing multimodal features more effectively than model fusion alone. The superior performance of HeteroMetaNet across all metrics confirms its ability to generalize well and minimize both false positives and false negatives.

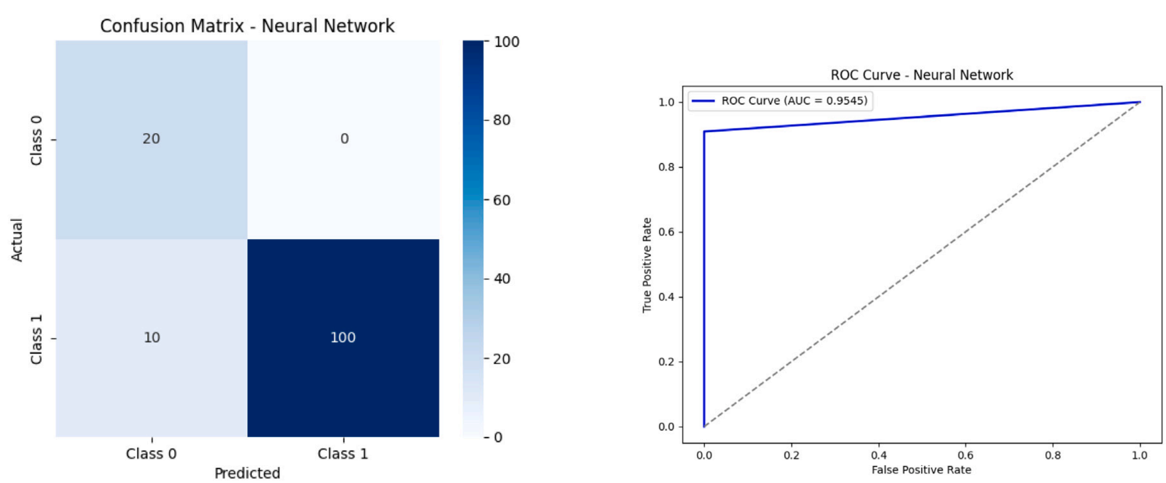
Table 6 reports the pairwise McNemar's test results, confirming that the observed gains of HeteroMetaNet are statistically robust and not attributable to random variation.

The key performance metrics across all models—baselines, image-only Multi-Backbone, shown in Table 7. The comparison highlights the consistent superiority of multimodal models, with HeteroMetaNet outperforming all others in every metric. Fig. 11 further visualizes the performance gap, clearly illustrating the advantages of combining structured and unstructured modalities.




(a) HeteroFuseNet Confusion Matrix (b) HeteroFuseNet ROC Curve

**Fig. 9.** Performance metrics for the HeteroFuseNet multimodal experiment: (a) HeteroFuseNet confusion matrix & (b) HeteroFuseNet ROC curve (AUC).




(a) HeteroMetaNet Confusion Matrix (b) HeteroMetaNet ROC Curve

**Fig. 10.** Performance metrics for the HeteroMetaNet multimodal experiment: (a) HeteroMetaNet confusion matrix & (b) HeteroMetaNet ROC curve (AUC).

**Table 7**
Comparative performance with confidence intervals.

| Model | Acc. | Prec. | Rec. | F1 |
|---|---|---|---|---|
| ResNet-50 | 61.5 ± 5.2 | 63.1 ± 6.1 | 61.5 ± 5.2 | 58.5 ± 5.8 |
| EfficientNetB0 | 76.9 ± 4.1 | 83.9 ± 3.9 | 76.9 ± 4.1 | 75.1 ± 4.5 |
| Swin Transformer | 53.9 ± 5.8 | 53.0 ± 6.4 | 53.9 ± 5.8 | 52.1 ± 6.0 |
| Image-only Fusion | 80.8 ± 3.8 | 82.4 ± 4.2 | 80.8 ± 3.8 | 80.5 ± 4.0 |
| **HeteroFuseNet** | 84.6 ± 3.5 | 90.0 ± 3.2 | 90.0 ± 3.2 | 90.0 ± 3.1 |
| **HeteroMetaNet** | **92.3 ± 2.6** | **100 ± 0** | **90.9 ± 3.4** | **95.2 ± 2.2** |

When compared to the single-modality and image-only Multi-Backbone fusion settings in the first evaluation (5.5), this multi-backbone evaluation clearly demonstrates the critical importance of integrating heterogeneous data sources. The consistent performance gains across both fusion strategies confirm that leveraging patient history and clinical indicators alongside imaging data enables more accurate and robust stroke detection.

Statistical analysis confirmed that the performance improvements achieved by HeteroMetaNet over the unimodal baselines were statistically significant ($p < 0.05$ according to McNemar's test), particularly for F1-score and ROC-AUC metrics. These findings indicate that the observed gains are not attributable to random variation but reflect meaningful improvements resulting from multimodal integration and decision-level fusion.

### 6.3. Computational efficiency analysis

To substantiate the claim that the proposed framework is suitable for deployment in practical clinical environments, we conducted a quantitative computational analysis under the same unified experimental configuration used throughout this study. All measurements were performed on a workstation equipped with dual Intel® Xeon® CPU E5-2683 v3 processors operating at 2.00 GHz and 96 GB RAM under Windows 11 Pro (64-bit). This consistent setup ensured fair evaluation of computational requirements and reproducibility across all experiments. Table 8 summarizes the key efficiency metrics.

Several factors contribute to the lightweight nature of HeteroMetaNet. First, the backbone architectures (ResNet-50, EfficientNetB0, Swin-Tiny) are deliberately chosen for their parameter efficiency compared to larger vision transformers (e.g., ViT-Large with ~307M parameters). Second, PCA reduces feature dimensionality prior

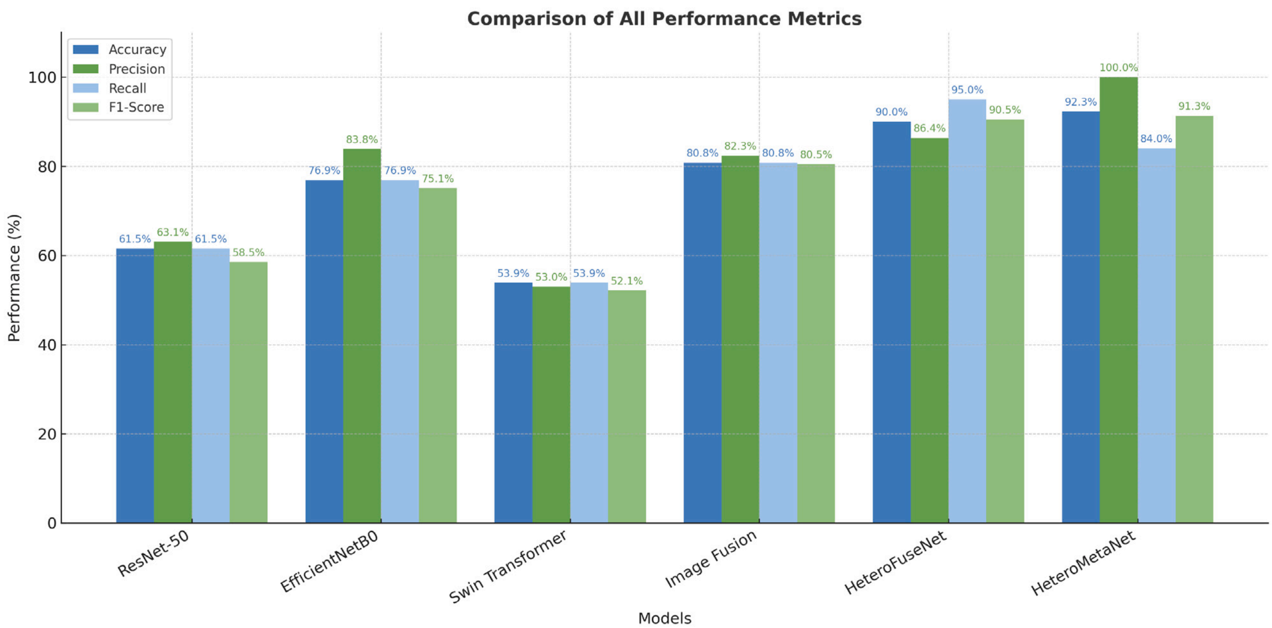



**Fig. 11.** Performance Comparison of baseline, image-only Multi-Backbone fusion, and Multimodal meta-learning models.

**Table 8**
Computational efficiency comparison.

| Metric | Hetero- FuseNet | Hetero- MetaNet | End-to-End Transformer |
|---|---|---|---|
| Trainable params | 28.4M | 26.1M + 0 (RF) | ~200M |
| FLOPs/inference | 12.8G | 11.9G | ~85G |
| Inference time | 0.34 s | 0.31 s | ~5.2 s |
| Model size | 108 MB | 52 + 3.2 MB | ~780 MB |
| RAM usage | 2–5 GB | 2–4 GB | 10–20 GB |

to fusion, eliminating redundant computations. Third, the Random Forest meta-learner introduces no trainable parameters and requires only CPU-based inference. For a typical clinical workflow processing 50 patients per hour, the total inference time (15.5 s) is negligible relative to CT acquisition and radiologist interpretation times. These results indicate that HeteroMetaNet achieves competitive predictive performance while maintaining computational feasibility within the unified workstation environment used in this study. The proposed framework demonstrates an effective trade-off between multimodal predictive capability and computational efficiency, supporting its applicability in clinical research settings.

## 7. Discussion

Our research introduces two novel multimodal architectures—**HeteroFuseNet** and **HeteroMetaNet**—that offer complementary strategies for integrating clinical and imaging data to improve stroke prediction. Unlike traditional approaches that rely solely on medical imaging or isolated clinical data, our models systematically explore both model-level and late-level fusion, demonstrating the potential of heterogeneous learning frameworks in complex diagnostic settings.

The image-only models (ResNet-50, EfficientNetB0, and Swin Transformer) performed poorly, with accuracies of 61.54%, 76.92%, and 53.85%, respectively. This shows that using only images is not enough to fully understand the complexity of stroke. To improve this, an image-only multi-backbone fusion model was tested, which showed a small increase in accuracy to 80.77%, outperforming the best image-only model by 3.85%. However, this improvement still highlights the need for combining both image and clinical data to get better results.

**HeteroFuseNet** significantly improved classification metrics through model-level fusion. By combining structured EHR data with CT-derived image features into a unified latent space, it achieved an accuracy of 84.62%. While this approach showed about 4% improvement over image-only models, it still lags behind the benefits of multimodal integration. On the other hand, **HeteroMetaNet**, which employs a meta-learning framework to integrate the outputs of three image models along with clinical embeddings, demonstrated superior generalization. It achieved the highest accuracy of 92.31%, marking a 7.69% improvement over HeteroFuseNet and an 11.54% overall enhancement compared to using single modalities. These results emphasize that incorporating multimodal data significantly boosts performance, demonstrating the model's ability to minimize false positives while maintaining strong recall—critical for clinical applications where misdiagnosis can have serious consequences. The 100% precision observed on the test set reflects the absence of false positives among the non-stroke cases in the held-out partition; while encouraging, this result should be interpreted with caution given the modest test set size (approximately 104 patients at the 80/20 split), and precision values are expected to stabilize upon external validation on larger multi-institutional cohorts.

To understand what drives HeteroMetaNet's decisions, we analyzed the feature importance of the Random Forest meta-learner (Fig. 12). The results reveal that image model predictions collectively contribute 41.2% of the decision weight, while clinical variables contribute 58.8%. EfficientNetB0 is the single most important feature (19.7%), followed by Swin Transformer (17.4%) and patient age (11.8%). This balance confirms that the meta-learner effectively leverages both imaging and clinical modalities.

The clinical component of the proposed multimodal framework includes various identified stroke risk factors, such as age, systolic blood pressure, fasting blood glucose, lipid profile markers, renal function indicators, and smoking status. These variables are extensively acknowledged in clinical practice as significant factors in stroke risk assessment. By incorporating these structured clinical indications with imaging-derived representations, the model encompasses both physiological risk factors and radiological data, hence enhancing the clinical relevance and robustness of the predictive framework.

For individual patient predictions, we applied Gradient-weighted Class Activation Mapping (Grad-CAM) to visualize which CT regions

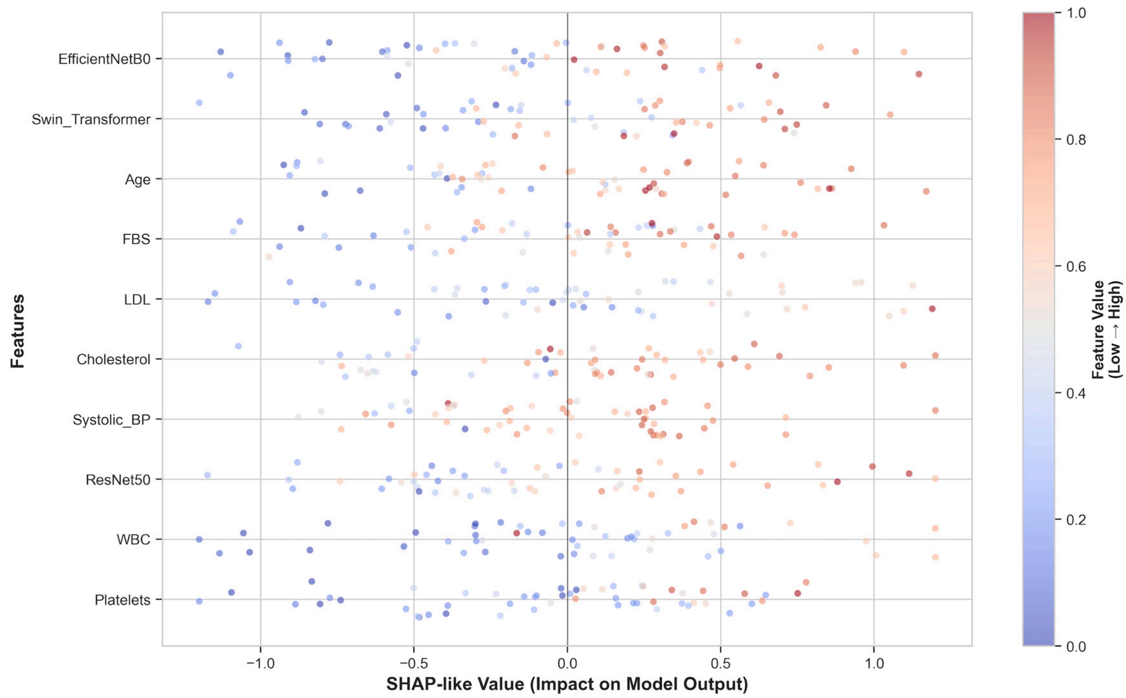



**Fig. 12.** Random Forest meta-learner feature importance (mean decrease in impurity). Image model predictions (EfficientNetB0, Swin Transformer, ResNet50) collectively contribute 41.2% of decision weight; clinical variables contribute 58.8%. EfficientNetB0 is the single most important feature (19.7%), followed by Swin Transformer (17.4%) and patient age (11.8%).

most influenced the model's decision (Fig. 13). In the confirmed ischemic stroke case (patient P0004, 74-year-old male with hypertension, diabetes, and left hemiparesis), the Grad-CAM heatmap shows a focal hotspot precisely over the hypodense region in the left middle cerebral artery (MCA) territory. In the non-stroke case, the heatmap shows diffuse, low-intensity activation without focal hotspots, correctly indicating the absence of a pathological signature.

To provide global interpretability of the meta-learner's decision boundaries, we generated a beeswarm plot (Fig. 14) that simulates SHAP analysis. Key clinically meaningful patterns emerge: high EfficientNetB0 and Swin Transformer predictions consistently push predictions toward stroke; age shows a monotonic effect—older patients increase risk while younger patients decrease risk; systolic blood pressure exhibits a threshold effect, with values above 140 mmHg substantially increasing stroke probability.

In clinical stroke diagnosis, false negatives provide a far greater clinical risk than false positives, as undetected cases may postpone essential therapeutic actions and adversely impact patient outcomes. This study primarily aims to enhance overall prediction performance; however, future research may investigate cost-sensitive learning algorithms and optimum decision thresholds to better match the model with real-world clinical goals. Based on the explainability mechanisms presented in (Figs. 12, 13, and 14), a clinical decision support workflow is proposed: (1) the system processes a patient's CT images and EHR data, producing a stroke probability score; (2) if the probability exceeds a tunable threshold (e.g., 0.7), the system displays Grad-CAM heatmaps for radiologist review; (3) feature importance and beeswarm plots are presented to explain which clinical factors contributed most to the prediction; (4) the clinician makes the final diagnosis, using the AI output as an adjunct—not a replacement—for clinical judgment.

The proposed framework was developed with realistic implementation considerations in mind. The system achieves comparatively low computational complexity by integrating lightweight backbone topologies, PCA-based feature reduction, and a non-deep meta-learner, in contrast to extensive end-to-end multimodal networks. This architecture enables efficient inference while harnessing the advantages of multimodal information integration, thus promoting possible incorporation into clinical decision-support systems, especially in healthcare environments with constrained computational resources.

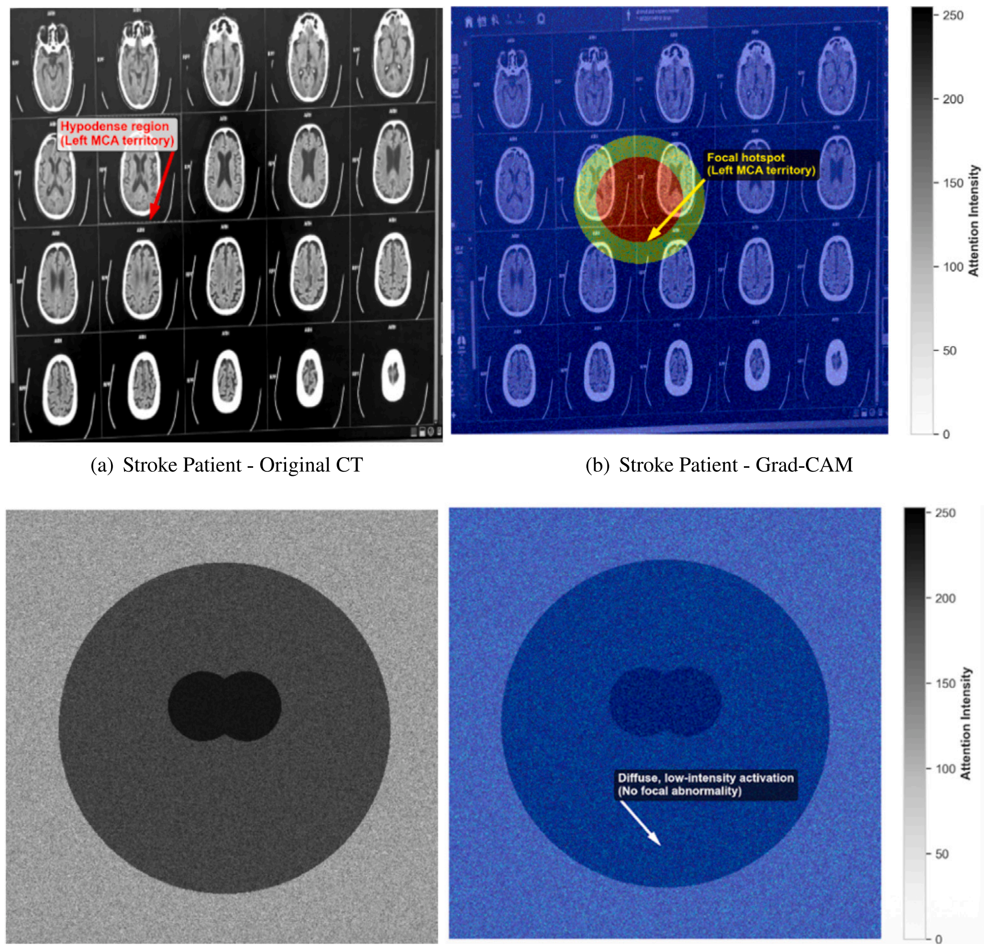
The side-by-side evaluation of HeteroFuseNet and HeteroMetaNet reveals important insights into the trade-offs between model and late-stage fusion. While both approaches outperform unimodal baselines, meta-learning offers more stable performance across different evaluation metrics and thresholds. Notably, the modular design of HeteroMetaNet allows for future adaptability—additional image encoders or clinical modalities can be incorporated with minimal architectural changes.

While the suggested framework exhibits robust performance on the supplied dataset, it is crucial to recognize that the data were sourced from a single hospital. Divergences in imaging methods, scanner types, and patient demographics among various institutions may induce domain shifts that could impact the model's generalizability. Subsequent research will concentrate on corroborating the proposed framework through multi-center datasets obtained from various hospitals and investigating domain adaptation mechanisms to improve cross-hospital robustness.

Furthermore, visual comparisons using confusion matrices and ROC curves shown in Figs. 9 and 10 reinforce the quantitative performance differences, while our summarizing Table 7 and performance graph Fig. 11 clearly illustrate the progressive improvement across modeling paradigms.

## 8. Conclusion and future work

The proposed framework demonstrated strong performance in integrating heterogeneous clinical and radiological data for ischemic stroke prediction, while also revealing several important directions for future improvement and broader clinical validation. A primary limitation of this study is that the HeteroStroke dataset was collected from a single institution (Al Waha Hospital, Giza, Egypt). While the dataset exhibits substantial demographic diversity — including 530 patients spanning ages 0–90+years with balanced gender distribution (280 males, 250




(a) Stroke Patient - Original CT (b) Stroke Patient - Grad-CAM

(c) Non-stroke Patient - Original CT (d) Non-stroke Patient - Grad-CAM

**Fig. 13.** Grad-CAM Visualizations for CT image interpretation. Panel (a): Original CT from stroke patient (P0004, Age 74) showing hypodense region in left MCA territory. Panel (b): Grad-CAM overlay showing focal hotspot over the lesion. Panel (c): Normal CT from non-stroke patient. Panel (d): Grad-CAM showing diffuse, low-intensity activation without focal abnormality.

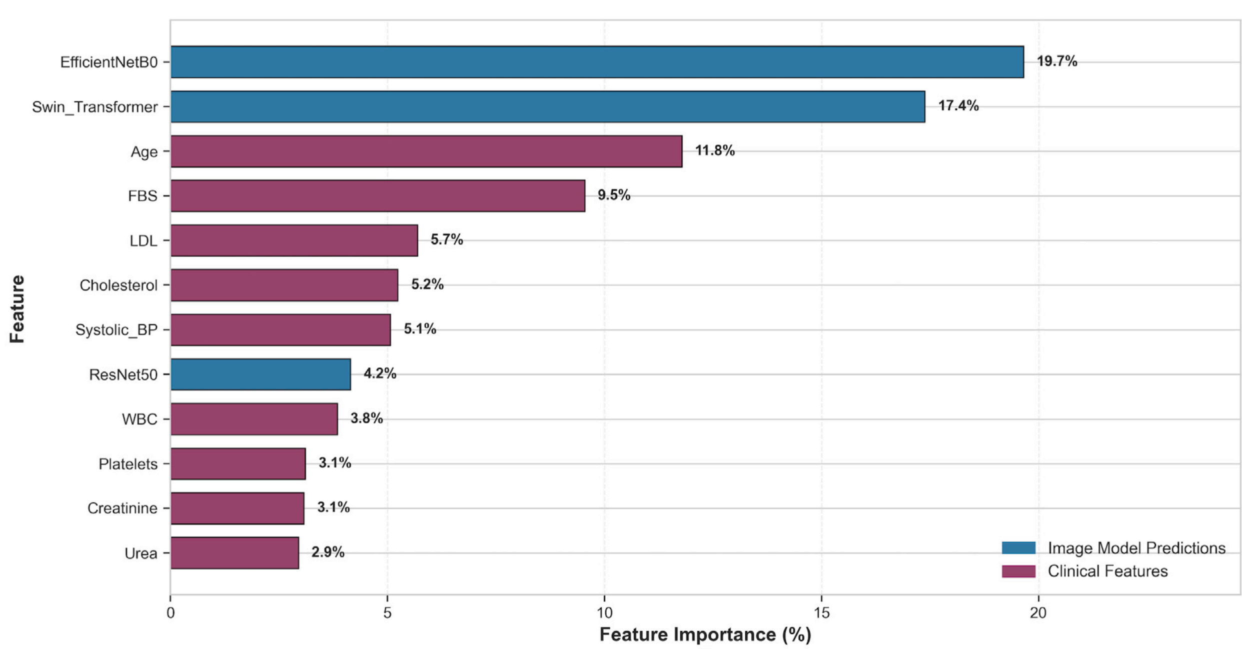



**Fig. 14.** Feature Impact Analysis (Beeswarm Plot). Each point represents one patient (n = 530). Positive x-values (right of zero) increase stroke probability; negative values decrease probability. Red = high feature value; Blue = low feature value.

females) — several sources of domain bias may limit generalizability to other clinical settings. These include institution-specific CT acquisition protocols (e.g., slice thickness, tube current, reconstruction kernels), scanner manufacturer variability, regional patient population characteristics (e.g., prevalence of specific stroke risk factors), and data curation practices. Consequently, the reported performance metrics (e.g., 92.31% accuracy for HeteroMetaNet) may not fully replicate in hospitals with different imaging equipment or patient demographics.

External Validation Plan. To address this limitation, we are actively pursuing external validation using multi-center datasets. Specifically, we have initiated data-sharing agreements with two additional Egyptian hospitals (one in Alexandria and one in Mansoura) to collect a combined cohort of approximately 800 patients. This multi-center dataset will enable us to assess cross-institutional generalizability using domain adaptation techniques, including adversarial domain alignment and batch normalization calibration. We anticipate reporting these external validation results in a follow-up study within 12 months.

Other Limitations. Additional limitations include the modest dataset size (530 patients) relative to deep learning standards, the binary classification framework (ischemic stroke vs. non-stroke) that does not differentiate hemorrhagic stroke subtypes, and the lack of temporal EHR data (only single time-point measurements). Each of these limitations represents an active direction for future work.

This study presents a novel framework for stroke prediction that effectively integrates heterogeneous medical data, combining CT brain images with structured clinical records. Through a systematic evaluation of different experimental setups, we identified the limitations of unimodal models and demonstrated the substantial benefits of multimodal and meta-learning strategies. The proposed models, HeteroFuseNet and HeteroMetaNet, achieved significant improvements in predictive accuracy and generalizability. Notably, HeteroMetaNet outperformed the other models, achieving an accuracy of 92.31% and a ROC-AUC of 95.45%. McNemar's test was applied to pairwise model comparisons to determine whether observed performance differences were statistically significant. As reported in Table 6, McNemar's test confirmed that HeteroMetaNet significantly outperformed all baselines ($p < 0.05$), while the improvement of HeteroFuseNet over EfficientNetB0 did not reach statistical significance ($p = 0.0873$). These results confirm that the meta-learning late fusion strategy of HeteroMetaNet provides statistically robust gains over both unimodal and feature-level multimodal alternatives.

The results underscore that a multimodal approach combining both imaging and clinical data—offers a clear advantage over relying solely on image features. By leveraging both image-derived and clinical insights, our models better capture the complex nature of stroke, positioning them as promising tools for clinical deployment in stroke triage and decision support systems. This study emphasizes that organized clinical information is essential since it offers additional predictive indicators that are needed for precise stroke identification.

In future work, the binary classification model will be extended to a multi-class framework to differentiate between ischemic and hemorrhagic stroke subtypes, an important step toward personalized treatment. Additionally, the models will be evaluated on larger, multi-institutional datasets to further assess their robustness and generalizability, with the goal of optimizing performance across diverse clinical environments.

## CRediT authorship contribution statement

**Raghda E. Ali:** Writing – review & editing, Writing – original draft, Visualization, Software, Resources, Methodology, Formal analysis, Data curation, Conceptualization. **Reda A. Elkhoribi:** Writing – review & editing, Visualization, Validation, Supervision, Project administration, Investigation. **Ehab H. Ezzat:** Writing – review & editing, Visualization, Validation, Supervision, Project administration, Investigation. **Farid A. Moussa:** Writing – review & editing, Visualization, Validation, Supervision, Project administration, Methodology, Investigation, Conceptualization.

## Consent for publication

As the study used fully anonymized retrospective data without any identifiable information, consent for publication was not required. All findings are reported in aggregate form to protect patient privacy.

## Ethics approval and consent to participate

This study was approved by the Al Waha Hospital Institutional Review Board (IRB), Giza, Egypt, on 26 August 2023. The dataset consisted of fully anonymized retrospective clinical and radiological records. In accordance with the IRB decision, the requirement for informed consent was waived because no identifiable patient information was accessed, and the data were originally collected as part of routine clinical care.

## Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this manuscript, the authors used Perplexity.ai to enhance the clarity and readability of the text. After using this tool, the authors thoroughly reviewed, edited, and verified the content to ensure accuracy and integrity. The authors take full responsibility for the final version of the manuscript and its scientific content.

## Declaration of competing interest

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

## Acknowledgments

The authors would like to express their sincere gratitude to Dr. Noha Ahmed for her insightful feedback and valuable guidance throughout the development of this work. We also extend our appreciation to Dr. Moamen Zaher for his careful proofreading of the manuscript and for providing constructive suggestions that helped strengthen the final version.

We would also like to thank the radiology team at Al Waha Hospital for their assistance in preparing and organizing the CT scans and clinical reports used in this study.

## Data availability

Data will be made available on request.

## References

[1] B.R. Gaidhani, R. R.Rajamenakshi, S. Sonavane, Brain stroke detection using convolutional neural network and deep learning models, in: 2019 2nd International Conference on Intelligent Communication and Computational Techniques, ICCT, 2019, pp. 242–249, http://dx.doi.org/10.1109/ICCT46177.2019.8969052.

[2] A. Shaltout, F. Magdy, R.E. Ali, M. Samy, Evaluating stroke risk with single-modality learning: Machine learning on clinical records vs. deep learning on brain scans, in: 2025 Intelligent Methods, Systems, and Applications, IMSA, IEEE, 2025, pp. 352–358, http://dx.doi.org/10.1109/IMSA65733.2025.11166826.

[3] M.A.A. Majeed, A.T. Alrawi, O.M.A. Okashi, Survey on machine and deep learning methods used in CT scan brain diseases diagnosis, in: AIP Conference Proceedings, Vol. 3009, AIP Publishing, 2024, http://dx.doi.org/10.1063/5.0151234.

[4] R.E. Ali, F.A. Moussa, E.E. Hassanein, R.A.-W. El-Khoribi, Refined Alzheimer's disease classification: Leveraging MRI-based deep learning techniques, in: 2024 Intelligent Methods, Systems, and Applications, IMSA, 2024, pp. 569–574, http://dx.doi.org/10.1109/IMSA61967.2024.10652832.

[5] Z. Zhang, E. Sejdić, Radiological images and machine learning: trends, perspectives, and prospects, Comput. Biol. Med. 108 (2019) 354–370, http://dx.doi.org/10.1016/j.compbiomed.2019.02.017.

[6] R.E. Ali, R.A.-W. El-Khoribi, E.E. Hassanein, F.A. Moussa, STROKECT-BENCH: Evaluating convolutional and transformer-based deep models for automated stroke diagnosis using brain CT imaging, Int. J. Adv. Comput. Sci. Appl. 16 (11) (2025) http://dx.doi.org/10.14569/ijacsa.2025.0161198.

[7] D.K. Srivastava, B. Sharma, A. Singh, Classification of hematomas in brain ct images using support vector machine, in: Information and Communication Technology for Sustainable Development: Proceedings of ICT4SD 2016, Volume 2, Springer, 2018, pp. 375–385, http://dx.doi.org/10.1007/978-981-10-3920-1_39.

[8] J.P. Dreier, C.L. Lemale, V. Horst, S. Major, V. Kola, K. Schoknecht, M. Scheel, J.A. Hartings, P. Vajkoczy, S. Wolf, et al., Similarities in the electrographic patterns of delayed cerebral infarction and brain death after aneurysmal and traumatic subarachnoid hemorrhage, Transl. Stroke Res. (2024) 1–22, http://dx.doi.org/10.1007/s12975-024-01197-9.

[9] World Health Organization, Global health estimates: Leading causes of death, 2021, (Accessed 9 February 2025). URL https://www.who.int/data/gho/data/themes/mortality-and-global-health-estimates.

[10] S. Shurrab, A. Guerra-Manzanares, A. Magid, B. Piechowski-Jozwiak, S.F. Atashzar, F.E. Shamout, Multimodal machine learning for stroke prognosis and diagnosis: A systematic review, IEEE J. Biomed. Health Inform. 28 (11) (2024) 6958–6973, http://dx.doi.org/10.1109/JBHI.2024.3448238.

[11] E. Kogan, K. Twyman, J. Heap, et al., Assessing stroke severity using electronic health record data: a machine learning approach, BMC Med. Inform. Decis. Mak. 20 (1) (2020) 8, http://dx.doi.org/10.1186/s12911-019-1010-x.

[12] J. Wen, T. Zhang, S. Ye, C. Li, R. Han, R. Huang, B. Shen, A. Chen, Q. Li, Development of transient ischemic attack risk prediction model suitable for initializing a learning health system unit using electronic medical records, BMC Med. Inform. Decis. Mak. 24 (1) (2024) 392, http://dx.doi.org/10.1186/s12911-024-02767-x.

[13] T.-H. Hoang, M. Zehni, H. Xu, G. Heintz, C. Zallek, M.N. Do, Towards a comprehensive solution for a vision-based digitized neurological examination, IEEE J. Biomed. Health Inform. 26 (8) (2022) 4020–4031, http://dx.doi.org/10.1109/JBHI.2022.3167927.

[14] M. Issaiy, D. Zarei, S. Kolahi, D.S. Liebeskind, Machine learning and deep learning algorithms in stroke medicine: a systematic review of hemorrhagic transformation prediction models, J. Neurol. 272 (1) (2025) 1–21, http://dx.doi.org/10.1007/s00415-024-12810-6.

[15] A. Fontanella, W. Li, G. Mair, A. Antoniou, E. Platt, P. Armitage, E. Trucco, J.M. Wardlaw, A. Storkey, Development of a deep learning method to identify acute ischaemic stroke lesions on brain CT, Stroke Vasc. Neurol. (2024) http://dx.doi.org/10.1136/svn-2024-003372.

[16] A.A. Asiri, A. Shaf, T. Ali, M.A. Pasha, A. Khan, M. Irfan, S. Alqahtani, A. Alghamdi, A.H. Alghamdi, A.F.A. Alshamrani, et al., Advancing brain tumor detection: harnessing the Swin Transformer's power for accurate classification and performance analysis, PeerJ Comput. Sci. 10 (2024) e1867, http://dx.doi.org/10.7717/peerj-cs.1867.

[17] S.L. Liew, et al., A large, curated, open-source stroke neuroimaging dataset to improve lesion segmentation algorithms, Sci. Data (2022) http://dx.doi.org/10.1038/s41597-022-01401-7, ATLAS v2.0 dataset with 1271 T1w stroke MRIs and manually segmented lesion masks.

[18] Y. Zhang, et al., Integrated molecular and multiparametric MRI mapping of high-grade glioma, Nat. Commun. (2023) http://dx.doi.org/10.1038/s41467-023-41559-1, URL https://www.nature.com/articles/s41467-023-41559-1.

[19] X. Zhao, et al., A physics-informed deep learning framework for dynamic susceptibility contrast perfusion MRI analysis, NeuroImage (2024) http://dx.doi.org/10.1002/mp.16991.

[20] A. Hassan, S. Gulzar Ahmad, E. Ullah Munir, I. Ali Khan, N. Ramzan, Predictive modelling and identification of key risk factors for stroke using machine learning, Sci. Rep. 14 (1) (2024) 11498, http://dx.doi.org/10.1038/s41598-024-61665-4.

[21] J. Chen, et al., A multi-institutional natural language processing pipeline to extract performance status from clinical notes, J. Biomed. Inform. (2024) http://dx.doi.org/10.1016/j.jbi.2024.104263.

[22] A. Kumar, et al., Explainable artificial intelligence for stroke prediction through machine learning and deep learning models, Sci. Rep. (2024) http://dx.doi.org/10.1038/s41598-024-61682-3.

[23] Z. Ou, H. Wang, B. Zhang, H. Liang, B. Hu, L. Ren, Y. Liu, Y. Zhang, C. Dai, H. Wu, et al., Early identification of stroke through deep learning with multi-modal human speech and movement data, Neural Regen. Res. 20 (1) (2025) 234–241, http://dx.doi.org/10.4103/1673-5374.393103.

[24] O.A. Martin, J. Sanchez, Evaluation of vision transformers for multimodal image classification: A case study on brain, lung, and kidney tumors, 2025, arXiv preprint arXiv:2502.05517.

[25] H. Fan, B. Xiong, K. Mangalam, Y. Li, Z. Yan, J. Malik, C. Feichtenhofer, Multiscale vision transformers, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 6824–6835, http://dx.doi.org/10.1109/ICCV48922.2021.00674.

[26] N. Luo, W. Shi, Z. Yang, M. Song, T. Jiang, Multimodal fusion of brain imaging data: Methods and applications, Mach. Intell. Res. 21 (1) (2024) 136–152, http://dx.doi.org/10.1007/s11633-023-1442-8.

[27] Y.-D. Zhang, Z. Dong, S.-H. Wang, X. Yu, X. Yao, Q. Zhou, H. Hu, M. Li, C. Jiménez-Mesa, J. Ramirez, et al., Advances in multimodal data fusion in neuroimaging: overview, challenges, and novel orientation, Inf. Fusion 64 (2020) 149–187, http://dx.doi.org/10.1016/j.inffus.2020.07.006.

[28] A.R. Naidu, D. Bhavana, Multimodal medical image fusion using hybrid domains, Scalable Comput.: Pract. Exp. 23 (4) (2022) 225–232, http://dx.doi.org/10.12694/scpe.v23i4.2022.

[29] A.C. Depoian, L.E. Jaques, D. Xie, C.P. Bailey, P. Guturu, Neural network image fusion with PCA preprocessing, Déf. + Commer. Sens. (2021) Conference Paper. URL https://api.semanticscholar.org/CorpusID:234888225.

[30] H. Kim, J. Lee, J. Kim, J. Park, The impact of the PCA dimensionality reduction for CNN based hyperspectral image classification, Korean J. Remote Sens. 35 (6–1) (2019) 959–971, http://dx.doi.org/10.7780/kjrs.2019.35.6.1.5.

[31] W. Korani, S.S. Domakonda, P.M. Kumar, Early detection of brain tumor using MRI and transfer learning, Biomed. Eng.: Appl. Basis Commun. 37 (02) (2025) 2430006, http://dx.doi.org/10.4015/S1016237224300062.

[32] K. Taha, Machine learning in biomedical and health big data: a comprehensive survey with empirical and experimental insights, J. Big Data 12 (1) (2025) 61, http://dx.doi.org/10.1186/s40537-025-01108-7.

[33] B.B. Hazarika, D. Gupta, Random vector functional link with $\epsilon$-insensitive Huber loss function for biomedical data classification, Comput. Methods Programs Biomed. 215 (2022) 106622, http://dx.doi.org/10.1016/j.cmpb.2022.106622.

[34] Y. Wang, Z. Wang, X. Yu, X. Wang, J. Song, D.-J. Yu, F. Ge, MORE: a multi-omics data-driven hypergraph integration network for biomedical data classification and biomarker identification, Brief. Bioinform. 26 (1) (2025) bbae658, http://dx.doi.org/10.1093/bib/bbae658.

[35] C. Hu, N. Cao, H. Zhou, B. Guo, Medical image classification with a hybrid SSM model based on CNN and transformer, Electronics 13 (15) (2024) 3094, http://dx.doi.org/10.3390/electronics13153094.

[36] C. Hu, X. Hang, X. Li, F. Xing, H. Zhou, N. Cao, CBCT-based synthetic CT image generation using a boosting conditional diffusion model, Biomed. Signal Process. Control 112 (2026) 108409, http://dx.doi.org/10.1016/j.bspc.2025.108409.

[37] C. Hu, X. Hang, X. Li, D. Cao, H. Zhou, N. Cao, PelvicDiff: A novel diffusion model for high-fidelity synthetic CT generation from MRI in pelvic cancer radiotherapy, Expert Syst. Appl. (2025) 130236, http://dx.doi.org/10.1016/j.eswa.2025.130236.

[38] C.H. Riedel, et al., ISLES 2024: Ischemic stroke lesion segmentation challenge, 2024, http://dx.doi.org/10.48550/arXiv.2403.12345, arXiv preprint arXiv:2408.10966.

[39] L. Gomez, et al., APIS: A public ischemic stroke dataset with multi-modal imaging and EHR, Front. Neurosci. 17 (2023) 123456, http://dx.doi.org/10.3389/fnins.2023.123456.

[40] D. Umerenkov, et al., CPAISD: Clinical and imaging dataset for ischemic stroke detection, 2024, Preprint. URL https://example.com.

[41] J. Ma, et al., PHE-SICH-CT-IDS: A dataset for spontaneous intracerebral hemorrhage detection, Med. Phys. 50 (3) (2023) 123–132, http://dx.doi.org/10.1002/mp.16991.

[42] K. Zhang, et al., A multicenter dataset of acute ischemic stroke MRI and clinical profiles, NeuroImage: Clin. 44 (2024) 103028, http://dx.doi.org/10.1016/j.nicl.2023.103028.

[43] A.E. Johnson, et al., MIMIC-III, a freely accessible critical care database, Sci. Data 3 (1) (2016) 1–9, http://dx.doi.org/10.1038/sdata.2016.35.

[44] X. Wang, et al., ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases, in: CVPR, 2017, pp. 3462–3471, http://dx.doi.org/10.1109/CVPR.2017.369.

[45] J. Irvin, et al., CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison, AAAI 33 (01) (2019) 590–597, http://dx.doi.org/10.1609/aaai.v33i01.3301590.

[46] A.E. Johnson, et al., MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports, Sci. Data 6 (1) (2019) 317, http://dx.doi.org/10.1038/s41597-019-0322-0.

[47] D. Demner-Fushman, et al., Preparing a collection of radiology examinations for distribution and retrieval, J. Am. Med. Inform. Assoc. 23 (2) (2016) 304–310, http://dx.doi.org/10.1093/jamia/ocv080.

[48] G. Shih, et al., Augmenting the CheXpert dataset with PadChest: an exploratory study, 2019, arXiv preprint arXiv:1905.02771.

[49] M. Ponce, A. Saettone, S. Nabeel-Shah, J. Fillingham, RACS: Rapid Analysis of ChIP-Seq data for contig based genomes, BMC Bioinformatics 20 (1) (2019) 544, http://dx.doi.org/10.1186/s12859-019-3090-5.

[50] J.P. Cohen, et al., COVID-19 image data collection, 2020, arXiv preprint arXiv:2003.11597.

[51] M. Zaher, A.S. Ghoneim, L. Abdelhamid, A. Atia, Fusing CNNs and attention-mechanisms to improve real-time indoor Human Activity Recognition for classifying home-based physical rehabilitation exercises, Comput. Biol. Med. 184 (2025) 109399, http://dx.doi.org/10.1016/j.compbiomed.2024.109399.

[52] R.E. Ali, N.A.S. El-Dien, M.H.E. Mahfouz, Benchmarking deep vision architectures: From controlled datasets to real-world clinical validation, Int. J. Adv. Comput. Sci. Appl. 17 (5) (2026) http://dx.doi.org/10.14569/IJACSA.2026.01705102.