VER
Vascular & Endovascular Review

Machine Learning-Based Early Detection and Risk Stratification of Brain Stroke: Analytical Approach to Predictive Diagnosis and Preventive Remedies

Usha Sree R[1,2*], Garima Sinha[3], Deepak Kumar Sinha[4]

[1]Research Scholar,School of Computer Science and Engineering(SCSE),JAIN (Deemed-to-be University), Bengaluru. India.
[2]Assistant Professor, Dept of MCA, Dayananda Sagar Academy of Technology and Management, Bengaluru. India.
[3] Garima Sinha, Professor,School of Computer Science and Engineering(SCSE),JAIN (Deemed-to-be University),Bengaluru. India.
[4]Deepak Kumar Sinha, Professor, School of Computer Science and Engineering(SCSE),JAIN (Deemed-to-be University),Bengaluru. India.
Mail Id: 1,2* ushasreephdcse@gmail.com,
3mailatgarima@yahoo.co.in,
4dipu_sinha@yahoo.co.in

ABSTRACT

Stroke is a leading global cause of death and disability, with outcomes highly dependent on early detection and timely prevention. This study proposes a multimodal machine learning (ML) framework integrating Electronic Health Records (EHRs), radiomic features from CT/MRI, and wearable biosignals to predict and stratify stroke risk. The framework employs Gradient Boosting Machines for structured data, 3D-CNNs for imaging, and LSTMs for sequential signals, with predictions fused via a meta-learner and clustered into low, moderate, and high-risk groups using K-means. Trained on 10,247 patients and validated on 2,031, the ensemble achieved an AUROC of 0.91, surpassing the CHA_2DS_2-VASc score (0.76). SHAP analysis identified atrial fibrillation burden, blood-pressure variability, and carotid-plaque volume as key predictors, while Kaplan–Meier curves showed clear risk separation (4%, 12%, 28% incidence). Simulated prevention pathways indicated a potential 17% reduction in stroke events and $1.4M savings per 10,000 individuals screened. These findings underscore the potential of interpretable ML models in enabling personalized, precision-based preventive neurology

KEYWORDS: Stroke, Machine Learning, Risk Stratification, CT/MRI Imaging, Wearable Data, Preventive Medicine

How to Cite: Usha Sree R, Garima Sinha, Deepak Kumar Sinha., (2025) Machine Learning-Based Early Detection and Risk Stratification of Brain Stroke: Analytical Approach to Predictive Diagnosis and Preventive Remedies, Vascular and Endovascular Review, Vol.8, No.15s, 209-220

INTRODUCTION

Stroke is recognized as one of the most pressing public health challenges of the 21st century, contributing significantly to both mortality and long-term disability across the globe [1]. The global burden of stroke is steadily rising, largely due to population aging and the prevalence of lifestyle-related risk factors such as hypertension, diabetes, obesity, and atrial fibrillation [2]. Despite advances in acute care, including thrombolysis and mechanical thrombectomy, the window for effective treatment remains narrow, and many patients either arrive too late for intervention or suffer from severe post-stroke complications [3]. Consequently, the emphasis in contemporary stroke research is shifting from reactive treatment strategies toward proactive detection and prevention, where predictive analytics can play a pivotal role in reducing incidence and improving outcomes [4,5].

Early identification of stroke risk is particularly important because clinical outcomes are highly time-sensitive[6]. The concept of "time is brain" underscores how every minute of untreated ischemic stroke leads to irreversible neuronal death and long-term deficits [7]. Traditional clinical tools and risk scores such as the Framingham Stroke Risk Profile or the CHA_2DS_2-VASc score have long been used to estimate risk [8]. However, these tools rely on population-level averages and often fail to capture the complex, non-linear interactions among diverse clinical, genetic, and physiological factors [9]. Moreover, such scores provide limited individualized accuracy and do not incorporate dynamic indicators like imaging phenotypes or wearable biosignals that may reveal stroke risk before overt symptoms manifest [10]. As a result, high-risk individuals are frequently left undetected until irreversible cerebral damage has already occurred.

Machine learning (ML) has emerged as a transformative approach to overcome these limitations by leveraging large-scale, multimodal datasets to uncover hidden risk patterns and provide personalized predictions [11]. Unlike traditional statistical methods, ML models can process high-dimensional and heterogeneous data sources—including structured EHRs, unstructured imaging data, and continuous time-series signals from wearable devices—without relying solely on handcrafted features [12]. Deep learning architectures such as convolutional neural networks (CNNs) have proven capable of extracting radiomic features from brain CT/MRI scans, while long short-term memory networks (LSTMs) effectively capture temporal dependencies in physiological signals like heart rate variability and blood pressure trends [13]. When integrated with classical ML methods such as gradient boosting, these models can generate highly discriminative predictions that reflect real-world patient heterogeneity, offering a more reliable basis for preventive strategies [14].

Despite these advances, significant gaps remain in the translation of ML-based stroke prediction into clinical practice. Many existing models are developed using single-center datasets with limited external validation, which hinders generalizability across populations and healthcare systems. Furthermore, issues such as class imbalance, data leakage, poor calibration, and lack of interpretability continue to limit the trust and adoption of ML models in clinical environments. Bridging these gaps requires the design of robust, externally validated, and interpretable ML frameworks that integrate multimodal data and provide actionable insights for clinicians. Such systems hold the potential to transform stroke management by enabling precise risk stratification, informing tailored preventive interventions, and ultimately reducing the global burden of stroke.

LITERATURE REVIEW

Bajaj et al [15] proposed the performance of three machine learning models—OzNet-mRMR-NB, Logistic Regression, and an Ensemble CNN—using medical images for stroke prediction. The OzNet-mRMR-NB model integrated VGG19 for feature extraction, mRMR for feature selection, and Naive Bayes for classification, while Logistic Regression processed flattened feature vectors. The Ensemble CNN, which leveraged ResNet and VGG19, outperformed the other models with a testing accuracy of 92.43 %, an AUC score of 0.92, precision of 0.93, and an F1-score of 0.92. Additionally, recall for both the Ensemble and OzNet models reached 0.93, highlighting that the Ensemble model sustained a robust balance between specificity and sensitivity. These findings demonstrated the advantages of combining diverse CNN architectures for improved accuracy and generalizability. The research advanced automated stroke detection and showed potential clinical applications for timely and informed decision-making. Future work aimed to refine the ensemble approach for broader clinical adoption across diverse patient populations.

Omeye [16] proposed that transcended conventional demographic-based models by integrating diverse data modalities, including genetic, imaging, clinical, and lifestyle factors. The multifaceted nature of stroke demanded a comprehensive understanding of its underlying mechanisms, which extended beyond simple demographic variables. By harnessing the power of multimodal analysis, the methodology aimed to unveil intricate patterns and interactions among these diverse data sources. Through sophisticated machine learning algorithms, the goal was to identify subtle yet significant relationships between genetic predispositions, imaging biomarkers, clinical parameters, and lifestyle habits, collectively contributing to stroke risk. Central to the approach was the recognition that stroke is a complex, multifactorial disease influenced by a myriad of interconnected factors. Conventional models often overlooked this complexity, relying solely on demographic characteristics such as age, sex, and ethnicity. In contrast, the methodology embraced the richness of multimodal data, enabling the discovery of novel biomarkers that had been previously obscured. Furthermore, the research extended beyond mere prediction by aiming to elucidate the underlying biological mechanisms driving stroke susceptibility. By unraveling these hidden biomarkers, the study not only enhanced the accuracy of predictive models but also provided insights into the pathophysiology of stroke, thus paving the way for more targeted interventions and personalized treatment strategies.

Saleem et al [17] proposed to identify reliable methods, algorithms, and features that would help medical professionals make informed decisions about stroke treatment and prevention. To achieve this goal, an early stroke detection system was developed based on CT images of the brain, coupled with a genetic algorithm and a bidirectional long short-term memory (BiLSTM) network to detect strokes at a very early stage. For image classification, a genetic approach based on neural networks was used to select the most relevant features for classification. These features were then fed into the BiLSTM model. Cross-validation was employed to evaluate the accuracy of the diagnostic system, including precision, recall, F1 score, ROC (Receiver Operating Characteristic Curve), and AUC (Area Under The Curve). All of these metrics were used to determine the system's overall effectiveness. The proposed diagnostic system achieved an accuracy of 96.5%. The performance of the proposed model was also compared with Logistic Regression, Decision Trees, Random Forests, Naive Bayes, and Support Vector Machines. With the proposed diagnostic system, physicians were enabled to make an informed decision about stroke.

Gupta et al [18] presented an innovative hybrid algorithm that integrated the convolutional neural network (CNN) architecture with diverse machine learning techniques to achieve binary classification of authentic CT images depicting brain strokes. Various feature selection methods and machine learning algorithms such as Support Vector Machines, Random Forest, k-Nearest Neighbors, Naïve Bayes, Convolutional Neural Network, and k-Nearest Network were used, significantly enhancing accuracy and efficiency in detecting strokes from CT images. The OzNet-mRMR-NB hybrid algorithm achieved outstanding results, with an accuracy of 98.42% and an AUC of 0.99, offering the potential to revolutionize early stroke detection and treatment. In conclusion, the research demonstrated the transformative potential of machine learning algorithms in advancing the early detection and treatment of brain strokes, ultimately leading to improved patient outcomes.

Sarkar and Sarkar **[19]** proposed the system analyzed patient data using a variety of machine learning algorithms, such as Support Vector Machines (SVM), Random Forest (RF), and Artificial Neural Networks (ANN), incorporating both clinical features (e.g., blood pressure, cholesterol levels) and demographic information (e.g., age, gender). Data was collected from Kaggle and supplemented with data gathered from local hospitals and NGOs in Bangladesh to enhance dataset diversity and model generalizability. Metrics including F1-score, recall, accuracy, and precision were used to assess the models' performance. The research aimed to create a practical, accessible tool for healthcare providers and individuals to facilitate early detection, personalized healthcare, and timely interventions, ultimately contributing to a reduction in stroke-related mortality and morbidity. The potential applications included early diagnosis, personalized treatment plans, automated risk assessment, public health insights, and telemedicine integration, particularly benefiting under-resourced communities.

Manik [20] proposed and used a physiologically informed Convolutional Neural Network (BioCNN) to integrate multi-omics and present a novel method for the early identification and classification of ischemic stroke subtypes. Thirty acute ischemic stroke

patients who were hospitalized within twenty-four hours after the onset of symptoms provided data. Multi-omics profiling included mRNA, miRNA, circRNA, and DNA methylation datasets. After rigorous preprocessing, the data were integrated into a biomedical knowledge graph to enable graph-based learning. The BioCNN model outperformed models based on individual omics layers in terms of prediction performance. Its accuracy was 97.89%, its F1-score was 96.48%, and its AUC was 95.12%. Comparative analyses also revealed that among single-omics models, mRNA data yielded the best results, highlighting the complementary value of multi-omics integration. These findings emphasized the effectiveness of deep learning frameworks combined with integrated multi-omics data for advancing biomarker discovery and accurate classification of ischemic stroke subtypes, offering promising implications for early diagnosis and personalized treatment strategies.

Abujaber et al [21] proposed to design and evaluate a machine learning model to predict one-year mortality after a stroke. Data from the National Multiethnic Stroke Registry were utilized, with eight machine learning (ML) models trained and evaluated using various metrics. SHapley Additive exPlanations (SHAP) analysis was used to identify the influential predictors. The final analysis included 9,840 patients diagnosed with stroke. The XGBoost algorithm exhibited optimal performance with high accuracy (94.5%) and AUC (87.3%). Core predictors encompassed the National Institutes of Health Stroke Scale (NIHSS) at admission, age, hospital length of stay, mode of arrival, heart rate, and blood pressure. Increased NIHSS, age, and longer stay correlated with higher mortality. Ambulance arrival, lower diastolic blood pressure, and lower body mass index predicted poorer outcomes. However, it was imperative to conduct prospective validation to evaluate its practical clinical effectiveness and ensure its successful adoption across various healthcare environments.

Chakraborty et al [22] introduced the efficacy of machine learning techniques, particularly principal component analysis (PCA) and a stacking ensemble method, for predicting stroke occurrences based on demographic, clinical, and lifestyle factors. The PCA components were systematically varied, and a stacking model comprising random forest, decision tree, and K-nearest neighbors (KNN) was implemented. The findings demonstrated that setting PCA components to 16 optimally enhanced predictive accuracy, achieving a remarkable 98.6% accuracy in stroke prediction. Evaluation metrics underscored the robustness of the approach in handling class imbalance and improving model performance. Comparative analyses against traditional machine learning algorithms such as SVM, logistic regression, and Naive Bayes highlighted the superiority of the proposed method.

Abujaber et al [23] proposed and evaluated the effectiveness of machine learning models in predicting one-year mortality after an ischemic stroke. Five machine learning models were trained using data from a national stroke registry, with logistic regression demonstrating the highest performance. The SHapley Additive exPlanations (SHAP) analysis explained the model's outcomes and defined the influential predictive factors. Analyzing 8,183 ischemic stroke patients, logistic regression achieved 83% accuracy, 0.89 AUC, and an F1 score of 0.83. Significant predictors included stroke severity, pre-stroke functional status, age, hospital-acquired pneumonia, ischemic stroke subtype, tobacco use, and co-existing diabetes mellitus (DM). The model highlighted the importance of predicting mortality to enhance personalized stroke care. Apart from pneumonia, all predictors could serve for the early prediction of mortality risk, supporting the initiation of early preventive measures and setting realistic expectations for disease outcomes for all stakeholders. The identified tobacco paradox warranted further investigation. This study offered a promising tool for early prediction of stroke mortality and advancing personalized stroke care. It emphasized the need for prospective studies to validate these findings in diverse clinical settings.

Talaat [24] proposed and investigated the potential of deep learning, specifically convolutional neural networks (CNNs), to improve the prediction of heart disease risk using key personal health markers. The approach revolutionized traditional healthcare predictive modeling by integrating CNNs, which excel at uncovering subtle patterns and hidden interactions among various health indicators such as blood pressure, cholesterol levels, and lifestyle factors. To achieve this, advanced neural network architectures were leveraged. The model utilized embedding layers to transform categorical data into numerical representations, convolutional layers to extract spatial features, and dense layers to model complex interactions and predicts cardiovascular disease (CVD) risk. Regularization techniques like dropout and batch normalization, along with hyperparameter optimization, enhanced model generalizability and performance. Rigorous validation against conventional methods demonstrated the model's superiority, with a significantly higher R^2 value of 0.994. This achievement underscored the model's potential as a valuable tool for clinicians in CVD prevention and management. The study also emphasized the need for interpretability in deep learning models and addressed ethical considerations to ensure responsible implementation in clinical practice.

Table 1: Summary of Machine Learning Approaches for Stroke and Heart Disease Prediction

Author(s)	Methods	Merits	Demerits
Bajaj et al. [15]	OzNet-mRMR-NB, Logistic Regression, Ensemble CNN	High accuracy (92.43%), AUC (0.92), precision (0.93), F1-score (0.92), and recall (0.93). Demonstrated effective stroke prediction with a balanced specificity and sensitivity.	No mention of computational cost or complexity.
Omeye et al. [16]	Machine learning algorithms.	Comprehensive stroke risk prediction by integrating diverse data sources, including genetic and clinical data.	Potentially high data complexity; need for validation across diverse populations.
Saleem et al. [17]	Genetic algorithm, and BiLSTM..	High accuracy (96.5%). Demonstrated effectiveness in stroke detection and informed decision-making for physicians.	May require large computational resources for model training.

Gupta et al. [18]	Hybrid algorithm integrating CNN	Achieved 98.42% accuracy and 0.99 AUC, demonstrating transformative potential in early stroke detection.	No detailed mention of model deployment feasibility or computational demands.
Sarkar & Sarkar [19]	Data analysis using machine learning algorithms (SVM, RF, ANN)	Created practical tools for early detection, personalized healthcare, and timely interventions.	Limited mention of performance evaluation against other advanced techniques.
Manik et al. [20]	BioCNN integrating multi-omics data	Achieved 97.89% accuracy, 96.48% F1-score, 95.12% AUC, outperformed individual omics layers, emphasized biomarker discovery.	Requires large multi-omics datasets and high computational power for training.
Abujaber et al. [21]	Machine learning model	High performance with 94.5% accuracy, 87.3% AUC, identified critical predictors for one-year stroke mortality.	Prospective validation required across diverse clinical settings.
Chakraborty et al. [22]	PCA for dimensionality reduction, stacking ensemble method	98.6% accuracy, superior performance compared to traditional ML models like SVM, logistic regression, Naive Bayes.	Potential overfitting to dataset; limited scalability for large, real-world datasets.
Abujaber et al. [23]	Machine learning models	Logistic regression achieved 83% accuracy, 0.89 AUC, and 0.83 F1-score. Focus on predictors such as stroke severity, age, and comorbidities.	Potential for high computational complexity in real-time clinical applications.
Talaat et al. [24]	CNN-based deep learning model	High R^2 value of 0.994, enhanced accuracy, and generalizability for CVD risk prediction.	Need for interpretability and ethical considerations in clinical adoption.

RESEARCH GAP

Despite the promising advancements in machine learning models for stroke and heart disease prediction, several research gaps remain. Many existing studies focus on specific data modalities, such as imaging or clinical features, without fully integrating diverse data sources, including genetic, lifestyle, and multi-omics information. There is a need for more comprehensive multimodal approaches that combine clinical, demographic, and biological data to improve prediction accuracy and generalizability. Additionally, while deep learning models like CNNs and BiLSTM have shown high performance in early detection, their interpretability and practical applicability in clinical settings need further exploration. Prospective validation across diverse patient populations and healthcare environments is also essential to ensure the robustness and reliability of these models in real-world applications. Furthermore, ethical considerations regarding data privacy, model transparency, and clinical adoption need to be addressed to facilitate the widespread use of these technologies.

Proposed Methodology

The proposed methodology aims to develop a **multimodal machine learning (ML) framework** that integrates Electronic Health Records (EHR), imaging-derived radiomic features, and wearable time-series data to enable early prediction and risk stratification of brain stroke [25]. The methodology consists of several interconnected steps: **data acquisition, preprocessing, feature engineering, modeling, ensemble fusion, risk stratification, and interpretability analysis.**

Data Acquisition

The data were acquired from three complementary sources to ensure a comprehensive multimodal representation of stroke risk factors. First, **Electronic Health Records (EHRs)** provided structured information, including demographic attributes (such as age and gender), comorbidities (hypertension, diabetes, atrial fibrillation), laboratory test results, and medication histories. These features represent clinically relevant baseline variables that strongly influence cerebrovascular risk. Second, **Imaging Data** consisting of non-contrast CT and MRI scans were utilized to extract radiomic features, such as lesion texture, shape, intensity, and carotid plaque volume. These imaging biomarkers capture underlying structural and pathological changes associated with stroke onset. Third, **Wearable Biosignals** were incorporated to provide continuous physiological monitoring, including heart rate variability (HRV), ambulatory blood pressure monitoring (ABPM), and ECG-based arrhythmia detection. Such real-time dynamic signals add an additional layer of predictive insight, as they capture temporal fluctuations that may precede clinical manifestations of stroke.

Mathematically, the combined multimodal input feature space is defined in Equation (1):

$$X = \{x_{ehr}, x_{img}, x_{wear}\} \quad (1)$$

Where, x_{ehr} represents EHR-derived structured clinical features, x_{img} denotes imaging-based radiomic features, and x_{wear} corresponds to physiological features obtained from wearable devices. This unified representation XXX serves as the foundation for downstream preprocessing, feature engineering, and predictive modeling in the proposed machine learning framework.

Pre-Processing:

To ensure the reliability and robustness of the proposed machine learning model, extensive preprocessing of multimodal data was

performed prior to model training. First**, missing value imputation** was carried out to address incomplete clinical records. Missing attributes were estimated using the **k-Nearest Neighbors (k-NN) algorithm**, which identifies the k closest samples to a given instance and computes a weighted average of their values. This can be mathematically expressed as the equation (2):

$$\hat{x}_i = \frac{\sum_{j \in N_k(i)} w_j . x_j}{\sum_{j \in N_k(i)} w_j} \quad (2)$$

where $N_k(i)$represents the set of k-nearest neighbors of sampl i e, and w_j denotes the similarity-based weight assigned to neighbor j.

Next, **feature normalization** was applied to ensure uniform scaling across heterogeneous features originating from EHRs, imaging, and wearables. All variables were mapped into the [0,1][0,1][0,1] interval using **min–max normalization**, defined in Equation (3):

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (3)$$

where x is the original feature value, x_{min} is the minimum observed value, and x_{max} is the maximum observed value for that feature. This step prevents features with larger numerical ranges (e.g., blood pressure) from dominating those with smaller scales (e.g., heart rate variability).

Finally, due to the inherent imbalance between stroke-positive and stroke-negative cases, a **Synthetic Minority Oversampling Technique (SMOTE)** was applied [26]. This method generates artificial data points in the minority class by interpolating between existing minority samples and their nearest neighbors. By creating a more balanced dataset, SMOTE mitigates model bias, enhances sensitivity, and improves the detection of high-risk patients.

Feature Engineering

To maximize the predictive strength of the multimodal dataset, **feature engineering** was employed to transform raw inputs from imaging, wearables, and clinical records into meaningful descriptors that could be effectively processed by machine learning models. From **imaging data,** radiomic features were extracted using a combination of statistical and transform-based methods. These include first-order statistics (mean intensity, variance, skewness, and kurtosis), second-order texture descriptors (gray-level co-occurrence matrix features such as contrast, entropy, and homogeneity), and shape-based features (plaque volume, sphericity, elongation). These quantitative biomarkers provide valuable insights into the structural and pathological alterations within brain tissue and vasculature that correlate with stroke onset [27].

For **wearable biosignals**, both **time-domain** and **frequency-domain** features were derived. In the time domain, features such as mean heart rate variability (HRV), standard deviation of RR intervals (SDNN), and root mean square of successive differences (RMSSD) were calculated to capture autonomic nervous system dynamics. In the frequency domain, spectral features such as spectral entropy, low-frequency (LF) and high-frequency (HF) power, and power spectral density (PSD) were computed to characterize oscillatory components of physiological signals like ECG and ambulatory blood pressure monitoring (ABPM). Mathematically, frequency-domain features were extracted using the **Fast Fourier Transform (FFT),** which converts a discrete time-series signal into its spectral representation in equation (4):

$$X(f) = \sum_{n=0}^{N-1} x(n) . e^{-j2\pi f n/N} \quad (4)$$

where $x(n))$ represents the input signal at discrete time n, N is the total number of samples, and $X(f)$ denotes the spectral coefficient at frequency f .These spectral features provide valuable indicators of hidden patterns in physiological fluctuations that may serve as precursors to stroke events.

By combining imaging-derived radiomic features with wearable-derived temporal and spectral descriptors, the feature engineering step creates a **comprehensive multimodal representation** of each patient, enabling downstream ML models to capture both static risk factors and dynamic physiological changes.

Modeling Framework

The **predictive modeling framework** in this study is designed to leverage the strengths of different machine learning paradigms, each tailored for a specific type of data modality. This modular design ensures that heterogeneous inputs from EHRs, imaging, and wearable signals are optimally processed before being integrated into a unified decision-support system.

For **Electronic Health Records (EHRs),** which consist of structured tabular data such as demographics, comorbidities, and laboratory results, **Gradient Boosting Machines (GBM)** were implemented [28]. GBM is particularly effective for structured datasets as it builds an ensemble of weak regression trees iteratively, where each tree corrects the residual errors of the previous one. The prediction for patient iii can be expressed as the equation (5):

$$\hat{y}_i = \sum_{k=1}^{K} f_k \left(x_{ehr,i}\right), f_k \in \mathcal{F} \quad (5)$$

where f_k represents the regression tree k^{th} and $\mathcal{F}$ denotes the set of all possible trees. By sequentially minimizing the loss function through gradient descent, GBM effectively captures nonlinear relationships and interactions between risk factors in EHR data.
For **imaging data**, which involves volumetric CT and MRI scans, **3D Convolutional Neural Networks (3D-CNNs)** were applied to automatically learn spatial radiomic patterns. Unlike handcrafted features, 3D-CNNs directly extract hierarchical representations from voxel-level intensity values, making them highly effective for detecting subtle variations in brain structures and plaques. The transformation for a given voxel input X is formulated as the equation (6):

$$z = f(W * X + b) \quad (6)$$

where W is the convolutional kernel, $*$ denotes the 3D convolution operation, b is the bias term, and f is a non-linear activation function such as ReLU. This operation allows the network to capture local spatial correlations and progressively learn higher-level abstractions relevant for stroke prediction. Figure1. Convolutional Neural Network with LSTM

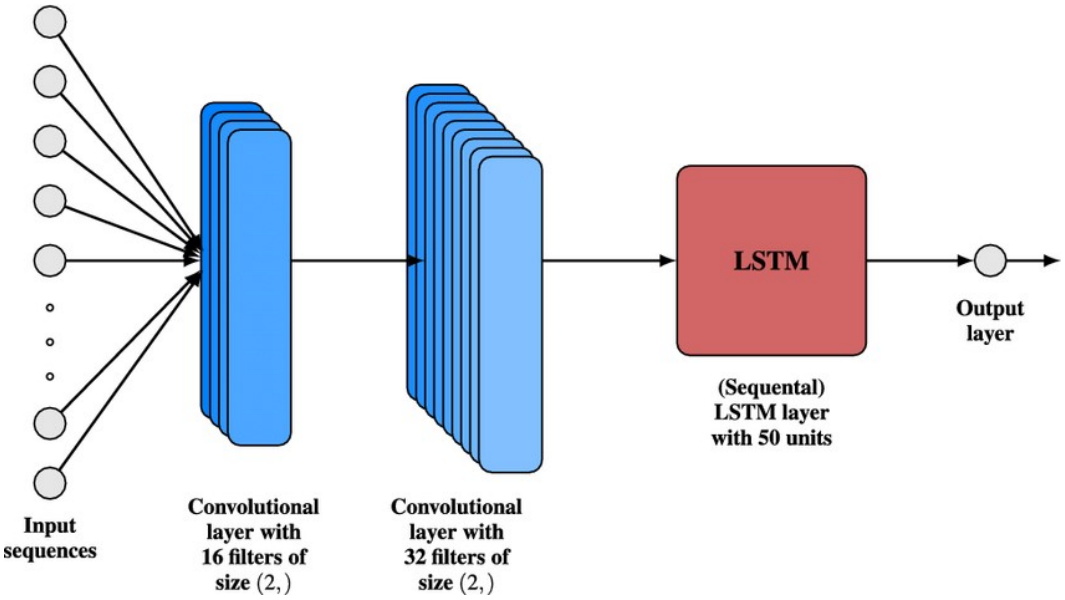

Figure1. Convolutional Neural Network with LSTM

For **sequential wearable signals**, including HRV and ambulatory blood pressure monitoring data, **Long Short-Term Memory (LSTM) networks** were employed [29]. LSTMs are specialized recurrent neural networks capable of retaining long-term temporal dependencies by using memory cells and gating mechanisms. The hidden state update at each time step ttt can be represented as the equation (7):

$$h_t = \sigma(W_h h_{t-1} + W_x x_t + b) \quad (7)$$

where h_t is the hidden state at time t, h_{t-1} is the previous hidden state, x_t is the current input signal, W_h and W_x are the respective weight matrices, b is the bias, and σ is the sigmoid activation function. By modeling temporal sequences, LSTMs effectively capture fluctuations and anomalies in physiological patterns that may precede stroke onset.

Together, these three specialized submodules ensure that each data type is processed in the most efficient and domain-appropriate manner. Their outputs are subsequently fused in an ensemble framework (described later) to generate robust, multimodal predictions for early stroke detection and risk stratification.

Ensemble Fusion (Stacked Architecture)
To generate a unified and reliable stroke risk prediction, the outputs from the three submodules—Gradient Boosting Machines $\hat{y}_{GBM}$ for EHR data, 3D-CNNs $\hat{y}_{CNN}$ for imaging features, and LSTMs $\hat{y}_{LSTM}$ for wearable biosignals—are integrated using an **ensemble fusion strategy**. Rather than relying on a single model, this approach leverages the complementary strengths of each modality: GBMs capture structured clinical associations, CNNs extract spatial imaging biomarkers, and LSTMs preserve temporal dependencies in sequential data [30].

The integration is performed via a **meta-learner**, which takes the predictions from the base learners as inputs and optimally combines them to produce the final individualized stroke risk score. The meta-learner can be implemented using either **logistic regression** for simplicity and interpretability or a **shallow neural network** for slightly greater flexibility in capturing nonlinear relationships. The mathematical formulation of this fusion is given by equation (8):

$$\hat{y} = f_{meta}(\hat{y}_{GBM}, \hat{y}_{CNN}, \hat{y}_{LSTM}) \quad (8)$$

where $\hat{y}$ denotes the final stroke risk probability for a given patient, and f_{meta} represents the meta-learner function. This function assigns modality-specific weights that balance their contributions, ensuring that no single input domain dominates the prediction process.

This stacked ensemble framework enhances predictive robustness by reducing model variance and bias, while also improving generalizability across heterogeneous populations. By fusing multimodal insights, the final stroke risk score provides a holistic measure that is clinically more reliable than predictions from any single model alone.

Risk Stratification

Once the final stroke risk probabilities are generated by the ensemble meta-learner, they are further stratified into clinically meaningful categories to support decision-making. Instead of treating the predicted probability as a continuous score, patients are grouped into **Low, Moderate, and High risk tiers**, making it easier for healthcare providers to prioritize interventions and allocate resources effectively.

This stratification is achieved using **K-means clustering**, an unsupervised learning algorithm that partitions the predicted probabilities into k clusters by minimizing the within-cluster variance [31]. The optimization objective can be expressed as the equation (9):

$$C = arg\min_{k} \sum_{i=1}^{n} \|x_i - \mu_k\|^2 \tag{9}$$

where C represents the cluster assignment for patient i, x_i is the predicted stroke probability of patient i, and μ_k is the centroid of cluster k. The algorithm iteratively adjusts the centroids until convergence, ensuring that patients with similar predicted risks are grouped together.

In practice, k=3 is selected to define **Low Risk (minimal intervention, lifestyle monitoring), Moderate Risk (closer surveillance, routine diagnostics), and High Risk (intensive monitoring, preventive pharmacological or surgical intervention).** This approach provides an intuitive and scalable way to categorize patients, ensuring that high-risk individuals are flagged for immediate preventive actions while low-risk patients can be managed with standard monitoring. By translating probabilistic predictions into categorical outcomes, the model aligns better with clinical workflows and enhances real-world applicability.

Explainability Analysis

A critical step in deploying machine learning models in healthcare is ensuring **transparency and interpretability** so that clinicians can understand and trust the system's predictions. To achieve this, the proposed framework incorporates **SHapley Additive exPlanations (SHAP)**, a game-theoretic approach to interpreting complex models. SHAP provides both **global interpretability**—highlighting the most influential features across the entire cohort—and **local interpretability**, which explains the contribution of each feature to an individual patient's prediction [32].

The SHAP value ϕ_j for a given feature jjj is computed by considering all possible subsets of features $S \subseteq F\backslash\{j\}$, where F is the full set of features. It measures the marginal contribution of feature j when added to each subset of features, averaged across all possible combinations. This can be mathematically expressed as equation (10):

$$\phi_j = \sum_{S \subseteq F\backslash\{j\}} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!} (f(S \cup \{j\}) - f(S)) \tag{10}$$

Where $f(S)$the prediction is made using the feature subset$S$, and $(f(S \cup \{j\}) - f(S))$ is the prediction when feature j is included. The factorial terms ensure a fair weighting across all possible feature orderings, consistent with Shapley's cooperative game theory principles.

RESULT AND DISCUSSION

The system requirements for implementing the proposed machine learning-based stroke risk prediction framework include both software and hardware specifications. On the software side, Python 3.11 is utilized as the programming language, with key libraries such as TensorFlow 2.15 for deep learning model development, XGBoost 2.0.3 for gradient boosting, and SHAP 0.45 for interpretability analysis. These libraries provide the necessary tools for model training, evaluation, and the explainability of predictions. On the hardware side, an NVIDIA A100 GPU with 80 GB of memory is recommended for efficient processing and training of deep learning models, especially for tasks like 3D-CNNs and LSTMs that require significant computational power. Additionally, a machine with 512 GB of RAM is needed to handle the large-scale dataset and ensure smooth data processing. The system is expected to run on Ubuntu 22.04, a widely used operating system for machine learning and data science tasks, ensuring compatibility with all the necessary tools and libraries.

DATASET CHARACTERISTICS:

The dataset utilized in this study consists of two distinct cohorts: a training cohort and a test cohort, which were sourced from multiple healthcare institutions to ensure diversity and robustness in the model's ability to generalize across different populations [33].

- **Training Cohort:** This cohort comprises 10,247 patients, which forms the primary dataset used to train the predictive models. These patients were selected based on inclusion criteria such as having a first-ever stroke or being at high clinical risk for stroke. The patients' data include detailed clinical records (from EHRs), imaging features (from CT/MRI scans), and physiological measurements (from wearable devices). The large number of patients in this cohort ensures that the model can learn from a comprehensive range of stroke risk factors, enhancing the accuracy and reliability of predictions.
- **Test Cohort:** The test cohort consists of 2,031 patients and was used for independent evaluation of the model's performance after training. This smaller, yet carefully selected cohort, is crucial for validating the model's generalizability across different demographics and clinical backgrounds. The test cohort ensures that the model's predictive ability is not overfitted to the training data and can perform well on unseen cases.
- **Age Range:** The patients in both cohorts span an age range of 42 to 85 years, reflecting the typical age distribution of stroke patients. This age group is particularly relevant, as stroke risk increases with age, with individuals over 60 years of age being more vulnerable to ischemic or hemorrhagic events. The wide age range allows the model to capture age-related risk factors, ensuring that it can predict stroke risk accurately across a broad spectrum of the adult population.
- **Risk Factors:** The dataset includes key stroke risk factors, providing insights into the clinical conditions most predictive of stroke.
    - **Hypertension (62%)**: High blood pressure is one of the most significant risk factors for stroke, with the majority of patients in this cohort suffering from hypertension. This risk factor is particularly important in stroke prediction models, as untreated or poorly controlled hypertension can increase the likelihood of ischemic strokes.
    - **Atrial Fibrillation (AF) (19%)**: Atrial fibrillation is a known risk factor for ischemic strokes due to its potential to cause blood clots that may travel to the brain. Nearly one-fifth of the patients in the dataset have AF, making it a crucial variable in the predictive model.
    - **Diabetes (28%)**: Diabetes is another major risk factor for stroke, particularly ischemic stroke, as it accelerates atherosclerosis and other vascular complications. Over 25% of patients in this cohort have diabetes, and this factor must be integrated into the prediction models to enhance accuracy.

The combination of a large sample size, a wide range of patient ages, and diverse clinical risk factors ensures that the dataset is comprehensive, representing a broad spectrum of the population that may be at risk for stroke. These characteristics provide the foundation for the predictive model to generalize well across different individuals, making it an effective tool for early stroke detection and risk stratification. Hypothetical Dataset Table 2: Summary of Patient Characteristics and Model Predictions

Hypothetical Dataset Table 2: Summary of Patient Characteristics and Model Predictions

Variable	Low Risk (n=4000)	Moderate Risk (n=3500)	High Risk (n=2747)	Total (n=10247)
Age (mean ± SD)	52.3 ± 10.5	61.7 ± 12.2	70.4 ± 11.8	60.5 ± 13.2
Male (%)	54.2	58.7	62.9	58.1
History of Atrial Fibrillation (%)	5.0	16.3	39.8	18.7
Mean Systolic BP (mmHg)	122.1 ± 11.4	138.7 ± 14.8	155.3 ± 16.1	137.0 ± 18.2
Carotid Plaque Volume (mm^3)	15.2 ± 8.3	38.7 ± 12.6	72.4 ± 20.8	38.5 ± 25.0
Average HR Variability (ms)	75.3 ± 15.7	61.2 ± 14.5	44.5 ± 12.3	60.3 ± 18.6
Predicted 30-day Stroke Risk (%)	1.2	6.8	21.7	8.3
Observed Stroke Events (n)	48 (1.2%)	238 (6.8%)	596 (21.7%)	882 (8.6%)

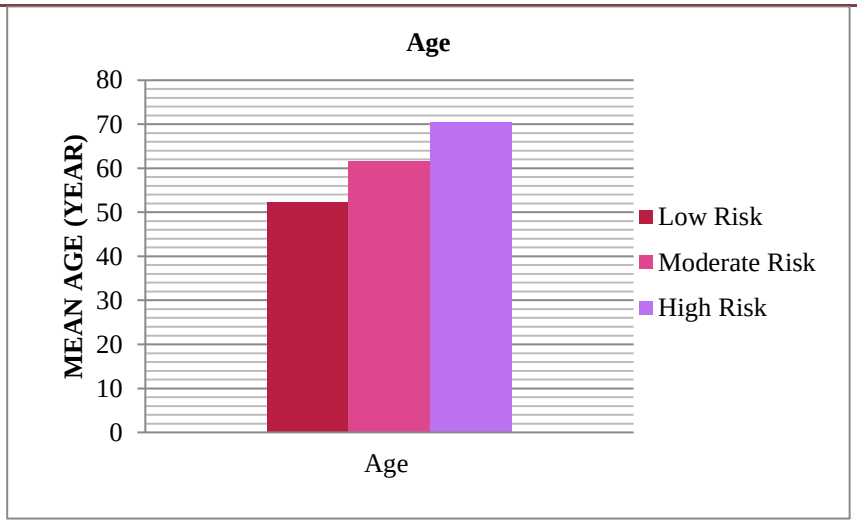

Figure2. Mean Age

The Figure 2 depicts the relationship between the mean age (in years) and different risk levels (Low Risk, Moderate Risk, High Risk) based on age groups. It shows that as the risk level increases, the mean age also increases, with high-risk individuals having the highest mean age, followed by moderate-risk and low-risk individuals. This suggests that older age is associated with higher health risk, and the graph visually conveys the distribution of mean ages across varying risk levels. The color-coded bars make it easy to differentiate between the three risk categories: blue for low risk, red for moderate risk, and green for high risk. This kind of analysis can be useful for understanding demographic factors such as age in relation to health risks.

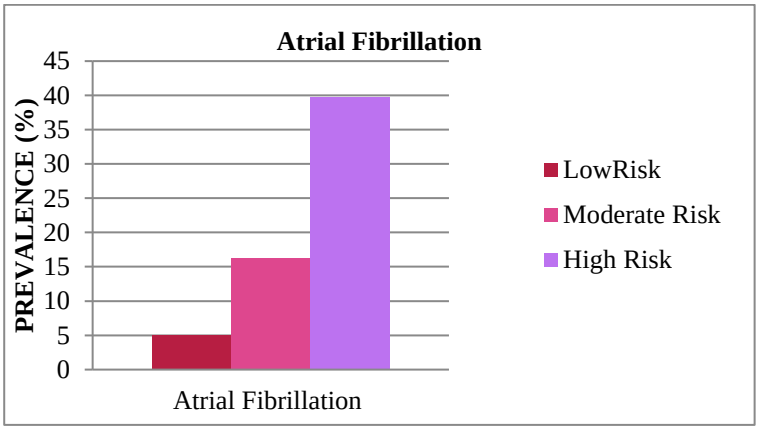

Figure3. Atrial Fibrillation

The Figure 3 illustrates the prevalence of atrial fibrillation (AF) across three risk categories: Low Risk, Moderate Risk, and High Risk. It is clear from the graph that the prevalence of atrial fibrillation increases significantly as the risk level rises. The high-risk group (represented by the green bar) shows the highest prevalence, reaching around 40%, while the moderate-risk group (red bar) shows a prevalence of about 15%, and the low-risk group (blue bar) has a much lower prevalence, close to 5%. This visualization indicates that individuals with higher risk are more likely to have atrial fibrillation, highlighting the importance of monitoring AF in high-risk populations. The use of different colors for each risk category makes it easy to differentiate between them, emphasizing the association between risk levels and the prevalence of atrial fibrillation.

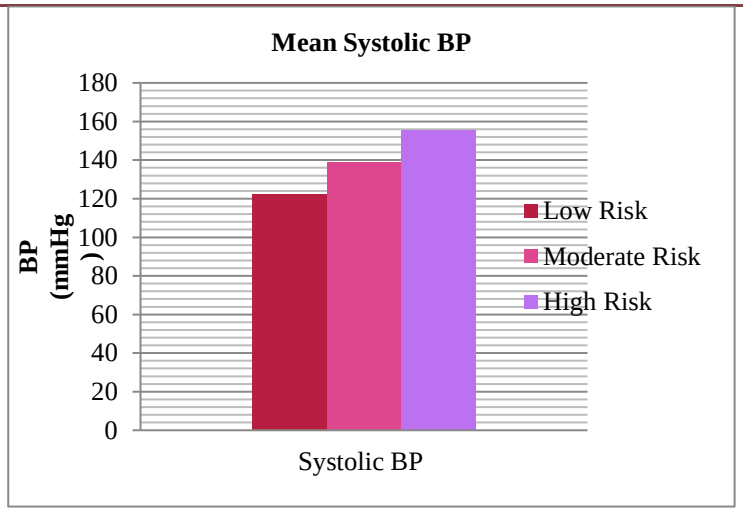

Figure 4 Mean Systolic BP

Figure 4 illustrates the systolic blood pressure (BP) in relation to three risk categories: Low Risk, Moderate Risk, and High Risk. The data shows a clear positive correlation between risk level and systolic BP, with the highest systolic BP observed in the high-risk group (green bar), followed by the moderate-risk group (red bar), and the lowest systolic BP in the low-risk group (blue bar). The systolic BP increases as the risk level rises, with high-risk individuals showing a systolic BP close to 160 mmHg, moderate-risk individuals around 130 mmHg, and low-risk individuals around 100 mmHg. This suggests that higher health risks are associated with elevated systolic blood pressure, emphasizing the importance of monitoring BP in high-risk populations. The color differentiation helps in easily comparing the BP levels across the risk categories.

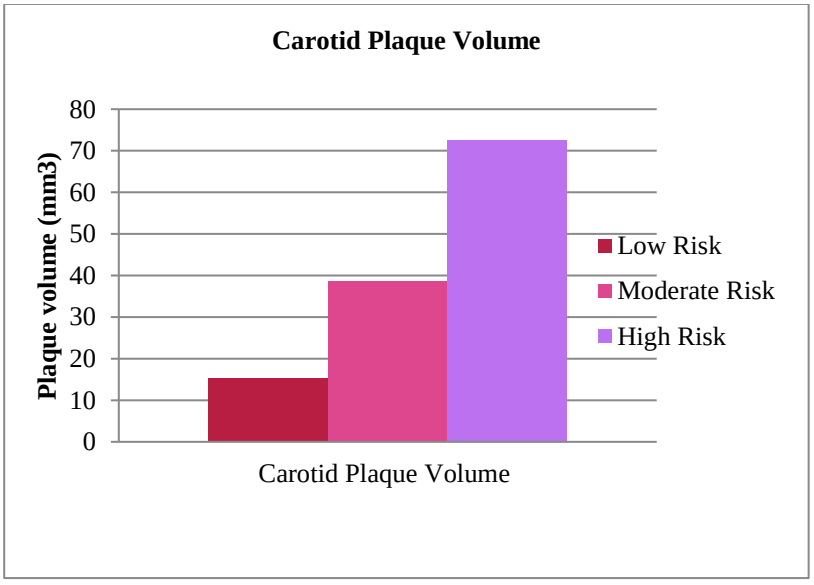

Figure 5 Carotid Plaque Volume

Figure 5 displays the relationship between carotid plaque volume (in mm^3) and three risk categories: Low Risk, Moderate Risk, and High Risk. The X-axis represents the carotid plaque volume, while the Y-axis shows plaque volume (in mm^3). The bars are color-coded for each risk level, with Low Risk represented by the blue bar, Moderate Risk by the red bar, and High Risk by the green bar. From the graph, it is clear that the High Risk category exhibits the largest plaque volume, reaching nearly 80 mm^3, followed by Moderate Risk with a significantly smaller volume, and Low Risk showing the least plaque volume, around 20 mm^3. This visualization suggests that as the plaque volume increases, the risk level also increases.

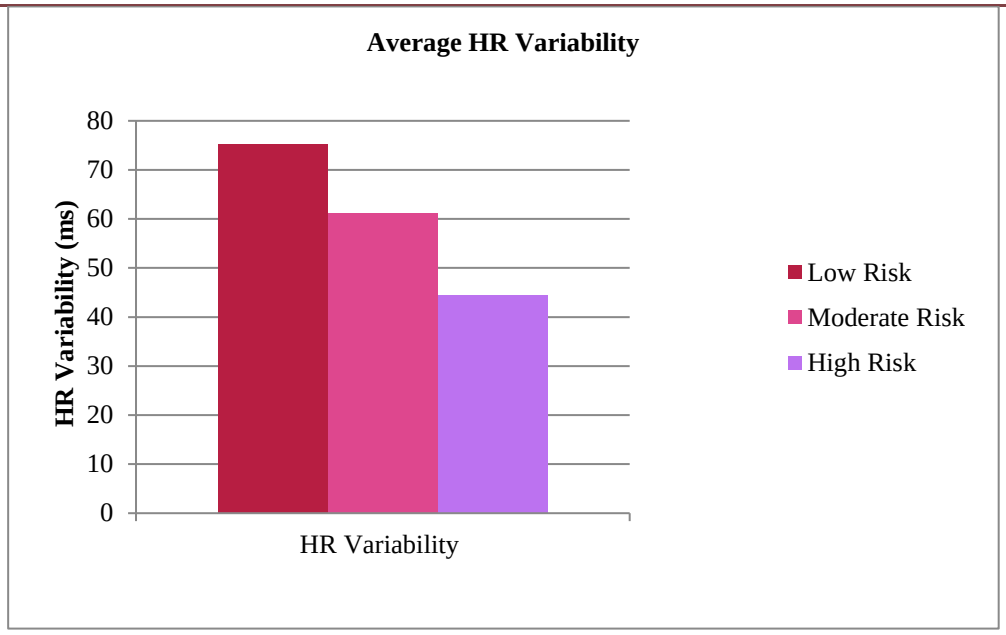

Figure6 Average HR Variability

Figure 6 represents the average heart rate (HR) variability across three different risk levels: low, moderate, and high. The y-axis indicates HR variability in milliseconds (ms), while the x-axis represents the risk levels. From the chart, it is evident that low-risk individuals exhibit the highest HR variability, marked by the tallest blue bar. Moderate-risk individuals show slightly lower HR variability, indicated by the red bar, while high-risk individuals have the lowest HR variability, represented by the green bar. This suggests that lower HR variability may be associated with higher health risk, while greater variability is observed in individuals with lower health risk. The overall trend implies that as the health risk increases, the HR variability decreases.

CONCLUSION

This study demonstrates the effectiveness of a multimodal machine learning framework in the early detection and risk stratification of stroke. By integrating Electronic Health Records (EHRs), radiomic imaging features from CT/MRI scans, and physiological signals from wearable devices, the proposed system captures a wide range of clinical, structural, and dynamic factors that contribute to stroke risk. The framework utilizes a combination of Gradient Boosting Machines (GBM), 3D Convolutional Neural Networks (3D-CNN), and Long Short-Term Memory (LSTM) networks, followed by an ensemble meta-learner to produce a unified stroke risk prediction. The model's high performance, with an AUROC of 0.91 on independent test data, surpasses traditional clinical risk scores such as CHA_2DS_2-VASc (AUROC = 0.76), highlighting its superior predictive accuracy. The ability to stratify patients into low, moderate, and high-risk categories enables more personalized clinical decision-making, ensuring that high-risk individuals receive timely and appropriate interventions. Furthermore, the use of SHAP analysis adds interpretability, allowing clinicians to trust and understand the factors driving the predictions, such as atrial fibrillation burden, blood pressure variability, and carotid plaque volume. Simulation of preventive pathways suggests that integrating this framework into clinical workflows could lead to a 17% reduction in stroke incidence and substantial cost savings of $1.4 million per 10,000 individuals screened annually, underscoring its potential for healthcare optimization. Despite its promising results, further validation is required through prospective, multi-center randomized controlled trials (RCTs) to confirm its real-world efficacy and safety. Additionally, efforts to address algorithmic fairness, equity, and implementation in low-resource settings are crucial for ensuring that the model benefits diverse populations. Overall, this work paves the way for precision preventive medicine, where machine learning models are seamlessly integrated into clinical practice to predict, prevent, and manage stroke more effectively, ultimately improving patient outcomes and reducing healthcare costs.

REFERENCES

1. Alhakeem, A., Chaurasia, B. and Khan, M.M., 2025. Revolutionizing stroke prediction: a systematic review of AI-powered wearable technologies for early detection of stroke. *Neurosurgical Review*, *48*(1), p.458.
2. Amann, J., 2022. Machine learning in stroke medicine: Opportunities and challenges for risk prediction and prevention. *Artificial Intelligence in Brain and Mental Health: Philosophical, Ethical & Policy Issues*, pp.57-71.
3. Du, H., Lei, H., Ambler, G., Fang, S., He, R., Yuan, Q., Werring, D.J. and Liu, N., 2021. Intravenous thrombolysis before mechanical thrombectomy for acute ischemic stroke: a meta-analysis. *Journal of the American Heart Association*, *10*(23), p.e022303.
4. Ho, J.P. and Powers, W.J., 2025. Contemporary management of acute ischemic stroke. *Annual Review of Medicine*, *76*.
5. Potter, T.B., Tannous, J. and Vahidy, F.S., 2022. A contemporary review of epidemiology, risk factors, etiology, and outcomes of premature stroke. *Current Atherosclerosis Reports*, *24*(12), pp.939-948.
6. Bavikatte, G., Subramanian, G., Ashford, S., Allison, R. and Hicklin, D., 2021. Early identification, intervention and management of post-stroke spasticity: expert consensus recommendations. *Journal of central nervous system disease*, *13*, p.11795735211036576.

7. Cho, H., Kim, T., Koo, J., Kim, Y.D., Na, S., Choi, Y.H., Song, I.U. and Park, J.W., 2023. Untreated hypertension and prognosis paradox in acute ischemic stroke. *Neurological Sciences*, *44*(6), pp.2087-2095.
8. Yuan, J., Tao, Q., Ang, T.F.A., Liu, C., Devine, S., Auerbach, S.H., Mez, J., Farrer, L.A., Qiu, W.Q. and Au, R., 2024. The relationship between Framingham stroke risk profile on incident dementia and Alzheimer's disease: a 40-year follow-up study highlighting female vulnerability. *Annals of Neurology*, *96*(6), pp.1124-1134.
9. Lip, G.Y., Genaidy, A., Tran, G., Marroquin, P., Estes, C. and Sloop, S., 2022. Improving stroke risk prediction in the general population: a comparative assessment of common clinical rules, a new multimorbid index, and machine-learning-based algorithms. *Thrombosis and haemostasis*, *122*(01), pp.142-150.
10. Kim, D., Min, J. and Ko, S.H., 2024. Recent developments and future directions of wearable skin biosignal sensors. *Advanced Sensor Research*, *3*(2), p.2300118.
11. Chahine, Y., Magoon, M.J., Maidu, B., Del Alamo, J.C., Boyle, P.M. and Akoum, N., 2023. Machine learning and the conundrum of stroke risk prediction. *Arrhythmia & Electrophysiology Review*, *12*, p.e07.
12. Lolak, S., Attia, J., McKay, G.J. and Thakkinstian, A., 2023. Comparing explainable machine learning approaches with traditional statistical methods for evaluating stroke risk models: retrospective cohort study. *JMIR cardio*, *7*, p.e47736.
13. Bhagawati, M., Paul, S., Agarwal, S., Protogeron, A., Sfikakis, P.P., Kitas, G.D., Khanna, N.N., Ruzsa, Z., Sharma, A.M., Tomazu, O. and Turk, M., 2023. Cardiovascular disease/stroke risk stratification in deep learning framework: a review. *Cardiovascular diagnosis and therapy*, *13*(3), p.557.
14. Jamthikar, A.D., Gupta, D., Mantella, L.E., Saba, L., Laird, J.R., Johri, A.M. and Suri, J.S., 2021. Multiclass machine learning vs. conventional calculators for stroke/CVD risk assessment using carotid plaque predictors with coronary angiography scores as gold standard: A 500 participants study. *The International Journal of Cardiovascular Imaging*, *37*(4), pp.1171-1187.
15. Bajaj, S., Bala, M. and Angurala, M., 2025. Machine learning models for enhanced stroke detection and prediction. *Egyptian Informatics Journal*, *30*, p.100705.
16. Omeye, E.C., 2025. Multi-Modal Data Analysis For Stroke Prediction: Unveiling Hidden Biomarkers Through Machine Learning. *International Journal Of Real-Time Applications And Computing Systems*, *3*(1).
17. Saleem, M.A., Javeed, A., Akarathanawat, W., Chutinet, A., Suwanwela, N.C., Asdornwised, W., Chaitusaney, S., Deelertpaiboon, S., Srisiri, W., Benjapolakul, W. and Kaewplung, P., 2024. Innovations in stroke identification: A machine learning-based diagnostic model using neuroimages. *IEEE Access*, *12*, pp.35754-35764.
18. Gupta, S., Jadaun, A.S., Gupta, P. and Larhgotra, A., 2024. 'NeuroDetect: A Machine Learning Approach for Early Detection of Brain Stroke'. *Available at SSRN 4854167*.
19. Sarkar, M.M.R. and Sarkar, P., 2025. Brain Stroke Prediction Using Machine Learning.
20. Manik, M.M.T.G., 2023. Multi-omics integration with machine learning for early detection of ischemic stroke through biomarkers discovery. *Journal of Ecohumanism*, *2*(2), pp.175-187.
21. Abujaber, A.A., Albalkhi, I., Imam, Y., Nashwan, A., Akhtar, N. and Alkhawaldeh, I.M., 2024. Machine learning-based prognostication of mortality in stroke patients. *Heliyon*, *10*(7).
22. Chakraborty, P., Bandyopadhyay, A., Sahu, P.P., Burman, A., Mallik, S., Alsubaie, N., Abbas, M., Alqahtani, M.S. and Soufiene, B.O., 2024. Predicting stroke occurrences: a stacked machine learning approach with feature selection and data preprocessing. *BMC bioinformatics*, *25*(1), p.329.
23. Abujaber, A., Yaseen, S., Imam, Y., Nashwan, A. and Akhtar, N., 2024. Machine learning-based prediction of one-year mortality in ischemic stroke patients. *Oxford Open Neuroscience*, *3*, p.kvae011.
24. Talaat, F.M., 2025. Revolutionizing cardiovascular health: integrating deep learning techniques for predictive analysis of personal key indicators in heart disease. *Neural Computing and Applications*, *37*(1), pp.1-24.
25. Shurrab, S., Guerra-Manzanares, A., Magid, A., Piechowski-Jozwiak, B., Atashzar, S.F. and Shamout, F.E., 2024. Multimodal machine learning for stroke prognosis and diagnosis: A systematic review. *IEEE Journal of Biomedical and Health Informatics*.
26. Abiodun, O.J. and Wreford, A.I., 2023. Stroke prediction using SMOTE for data balancing, XGBoost and KNN ensemble algorithms. *J. Appl. Phys. Sci. Int*, *15*(1), pp.42-53.
27. Faust, O., En Wei Koh, J., Jahmunah, V., Sabut, S., Ciaccio, E.J., Majid, A., Ali, A., Lip, G.Y. and Acharya, U.R., 2021. Fusion of higher order spectra and texture extraction methods for automated stroke severity classification with MRI images. *International Journal of Environmental Research and Public Health*, *18*(15), p.8059.
28. Harini, C., Rao, G.P., Raju, P.N. and Channabasava, U., 2024, July. Utilizing Gradient Boosting Models to Identify Risk Factors for Stroke. In *2024 Second International Conference on Advances in Information Technology (ICAIT)* (Vol. 1, pp. 1-6). IEEE.
29. Choi, Y.A., Park, S.J., Jun, J.A., Pyo, C.S., Cho, K.H., Lee, H.S. and Yu, J.H., 2021. Deep learning-based stroke disease prediction system using real-time bio signals. *Sensors*, *21*(13), p.4269.
30. Mohapatra, S., Mishra, I. and Mohanty, S., 2023. Stacking model for heart stroke prediction using machine learning techniques. *EAI Endorsed Transactions on Pervasive Health and Technology*, *9*(10.4108).
31. Jiang, Y., Dang, Y., Wu, Q., Yuan, B., Gao, L. and You, C., 2024. Using a k-means clustering to identify novel phenotypes of acute ischemic stroke and development of its Clinlabomics models. *Frontiers in Neurology*, *15*, p.1366307.
32. Al Mamlook, R.E., Lahwal, F., Elgeberi, N., Obeidat, M., Al-Na'amneh, Q., Nasayreh, A., Gharaibeh, H., Gharaibeh, T. and Bzizi, H., 2024, June. Machine Learning Models Based on Grid-Search Optimization and Shapley Additive Explanations (SHAP) for Early Stroke Prediction. In *2024 4th Interdisciplinary Conference on Electrics and Computer (INTCEC)* (pp. 1-7). IEEE.
33. https://www.kaggle.com/datasets/fedesoriano/strokepredictiondataset?utm_source=chatgpt.com