Industrial Engineering Journal
ISSN: 0970-2555
Volume : 54, Issue 4, April : 2025

Implementation of AI-Based Brain Stroke Prediction Analysis System

Dr.V.SHANMUKHA RAO [1]
M.Tech,(Ph.D) Professor in
Computer Science Engineering
PSCMRCET
Vijayawada,INDIA.
dr.shanmukharao@pscmr.ac.in

M.ASHA LATHA[2]
Computer Science Engineering
PSCMRCET
Vijayawada,INDIA.
ashalatha119rk@gmail.com

K.DEVI PRIYANKA[3]
Computer Science Engineering
PSCMRCET
Vijayawada,INDIA.
kodalidevipriyanka@gmail.com

P.SASI BHUSHANI[4]
Computer Science Engineering
PSCMRCET
Vijayawada,INDIA.
p.sasibhushani@gmail.com

L.SIRI CHANDANA[5]
Computer Science Engineering
PSCMRCET
Vijayawada,INDIA.
sirichandanalingala@gmail.com

***Abstract*— Stroke prediction is crucial for early medical intervention and improved patient outcomes. Building on existing machine learning-based approaches, this study introduces an advanced ensemble methodology integrating Categorical Boosting and Stacking Classifier, achieving a remarkable 99% accuracy. The proposed system enhances model robustness by handling diverse datasets effectively while ensuring interpretability through SHAP-based explanations. To facilitate practical deployment, a Flask-based web application with user authentication is developed, ensuring secure access to stroke prediction insights. This extension significantly improves prediction accuracy, security, and usability, making it a valuable tool for early stroke detection in healthcare systems.**

Keywords— *Stroke Prediction, Machine Learning, Ensemble Methods, Categorical Boosting, Stacking Classifier, Explainable AI, SHAP, SMOTE, Flask Web Application, User Authentication.*

I. INTRODUCTION

Stroke is a life-threatening medical condition that occurs when blood flow to the brain is disrupted, leading to severe neurological impairments. It remains one of the leading causes of disability and death worldwide, necessitating early detection and timely intervention. Traditional diagnostic approaches rely heavily on clinical assessments and medical imaging, which may be time-consuming and inaccessible in resource-limited settings. With advancements in artificial intelligence (AI) and machine learning (ML), automated stroke prediction models have emerged as a promising solution to facilitate early diagnosis and improve patient outcomes.

Recent research has focused on developing machine learning models using algorithms like Decision Trees (DT), Stochastic Gradient Descent (SGD), K-Nearest Neighbors (KNN), Support Vector Machines (SVM), and XGBoost. However, these models face challenges such as imbalanced datasets, lack of interpretability, and dependency on specific hospital data, affecting their generalization capability. To address these issues, this study incorporates an ensemble-based approach, leveraging Categorical Boosting and Stacking Classifier to improve prediction accuracy. These advanced techniques enhance model robustness, allowing it to handle diverse datasets more effectively.

Furthermore, interpretability is a critical aspect of AI-driven healthcare applications. The study integrates SHAP (Shapley Additive Explanations) to provide transparency in model decision-making, helping medical professionals understand the factors influencing stroke predictions. Additionally, a Flask-based web application with user authentication is developed to ensure secure and easy access to automated stroke predictions. By combining advanced ensemble methods, data balancing techniques, and explainable AI, this research presents a comprehensive framework that enhances stroke prediction accuracy, security, and usability, paving the way for early medical intervention and improved healthcare outcomes.

II. LITERATURE SURVEY

a) Plasmatic retinol-binding protein 4 and glial fibrillary acidic protein as biomarkers to differentiate ischemic stroke and intracerebral hemorrhage:
https://pubmed.ncbi.nlm.nih.gov/26526443/

In order to enhance therapy and results, it is crucial to distinguish between acute ischaemic stroke and intracerebral haemorrhage (ICH) promptly. Our goal was to find new plasma biomarkers that could differentiate between different forms of stroke and use them in conjunction with current indicators for this treatment indication. Eleven of the plasma samples examined with 177 antibodies showed different levels of chemokines, growth factors, and angiogenic factors across the different forms of stroke ($p < 0.05$), compared to 36 patients with ischaemic stroke and 10 patients with ischaemic cardiomyopathy. Replicated results were seen for pigment epithelial-derived factor, apolipoprotein B100, and RPB4 out of five proteins examined in 16 patients with ischaemic stroke and 16 patients with ICH ($p < 0.05$). These proteins, GFAP, and the receptor for advanced glycation end product were studied in 38 samples from ischaemic stroke and 28 samples from ICH. At last, 100% subtype specificity was shown by RBP4 >61 µg/mL and GFAP <0.07 ng/mL. Also, GFAP <0.07 ng/mL and RBP4 >48.75 µg/mL were found to be independent predictors of stroke subtype, improving discrimination by 29% ($p < 0.0001$, according to multivariate logistic regression analysis; ORadj: 6.09 (1.3-28.57), $p = 0.02$). These two indicators have the potential to differentiate ICH from ischaemic stroke. For appropriate treatment, it is crucial to rapidly differentiate between ischaemic stroke and intracerebral haemorrhage. We found and confirmed that RBP4 and circulating GFAP are plasmatic biomarkers for subtyping hyperacute stroke. The use of these and other biomarkers has the potential to

improve patient outcomes by expediting the categorisation of stroke subtypes.

b) Stroke Risk Prediction with Machine Learning Techniques https://pmc.ncbi.nlm.nih.gov/articles/PMC9268898/

A stroke occurs when the brain's blood supply is abruptly cut off. Brain cells die and function is impaired due to a lack of blood, depending on the area impacted. Predicting a stroke and improving health can be achieved by early symptom identification. Several ML models are created and tested in this work to provide a reliable method for predicting the likelihood of stroke. Accuracy, precision, recall, AUC, and F-measure are some of the metrics used to evaluate the performance of the stacking method presented in this research. With a 98.9% AUC, 98% F-measure, precision, recall, and accuracy, stacking classification was shown to be the most effective method in the research.

c) Segnet: A deep convolutional encoder-decoder architecture for image segmentation

https://ieeexplore.ieee.org/document/7803544

The innovative and practical SegNet architecture employs deep fully convolutional neural networks for semantic pixel-wise segmentation. The 13 convolutional layers of VGG16 are topologically identical to an encoder network [1]. When it comes time to classify pixels by themselves, the decoder network takes feature maps from the encoder and transforms them into full-resolution input feature maps. SegNet's decoder uses an unconventional method to upsample its input feature map, which has lower resolution. After the encoder's max-pooling step, the decoder uses the pooling indices to do non-linear upsampling. By doing so, upsampling training is disabled. In order to generate dense feature maps, sparse upsampled maps are convolved using trainable filters. Notable alternatives to this design include DeepLab-LargeFOV [3], FCN [2], and DeconvNet [4]. The correlation between accuracy and memory performance in segmentation is seen by this comparison. The applications that drove SegNet were scene understanding. Memory and computational efficiency are its primary design goals when it comes to inference. With stochastic gradient descent, it is possible to train it end-to-end with a significantly smaller number of trainable parameters compared to earlier designs. SUN RGB-D indoor scene segmentation and road scene benchmarking used as controls for SegNet and other systems. When compared to competing architectures, SegNet performs better in terms of inference time and memory efficiency, according to these quantitative evaluations. Visit http://mi.eng.cam to see a live demonstration of our Caffe SegNet implementation.

d) Recent progress in semantic image segmentation: https://arxiv.org/abs/1809.10198

Medical and intelligent transportation use semantic image segmentation, a significant application of computer vision and image processing. Academics often use multiple benchmark datasets to evaluate algorithm performance. Research on semantic segmentation has been ongoing for quite some time. Segmentation has come a long way since the advent of Deep Neural Networks (DNNs). In this research, we categorise semantic image segmentation techniques as either old or modern DNN. We begin with a brief overview of the original method and available datasets for segmentation, and then we dive deep into eight different areas of recent DNN-based methods: pyramid methods, fully convolutional networks, upsample methods, FCN-CRF methods, dilated convolution, backbone network enhancements, multi-level feature and multi-stage methods. At last, this part comes to a close.

e) A fast learning algorithm for deep belief nets: https://pubmed.ncbi.nlm.nih.gov/16764513/

For inference to be easier in highly-connected belief networks with many hidden layers, we use "complementary priors" to reduce the impact of explaining away effects. If the first two layers of a directed belief network are an undirected associative memory, we may train the network one layer at a time using a greedy, fast technique that uses complimentary priors. A slower learning technique that uses a contrastive wake-sleep strategy to fine-tune weights is initiated by the quick, greedy approach. After some tweaking, a three-hidden-layer network can provide a robust generative model of handwritten digit images and label distribution. This generative model outperforms even the most advanced discriminative learning algorithms when it comes to classifying numbers. Substantial ravines reflect the low-dimensional manifolds on which the numbers lie in the top-level associative memory's free-energy landscape. Directed links make it easy to investigate these ravines and reveal the memory's intentions.

III. METHODOLOGY

The proposed methodology involves developing an advanced stroke prediction model using machine learning and ensemble techniques. Initially, data preprocessing is performed, including handling missing values and balancing the dataset using SMOTE to address class imbalance. Feature selection is conducted using Mutual Information Score, Chi-Square Score, and ANOVA to identify significant predictors. Various machine learning classifiers, including Decision Trees, SGD, KNN, SVM, and XGBoost, are evaluated for their predictive performance. To enhance accuracy, ensemble methods such as Categorical Boosting and Stacking Classifier are applied, with the latter achieving 99% accuracy. Explainable AI techniques like SHAP are integrated to interpret model decisions, ensuring transparency in predictions. A Flask-based web application is developed for real-time user access, incorporating authentication for secure and reliable stroke prediction deployment in healthcare settings.

A) Proposed System

The proposed system enhances stroke prediction by integrating advanced ensemble learning techniques, particularly Categorical Boosting and Stacking Classifier, to improve accuracy and model robustness. Traditional machine learning models often struggle with imbalanced datasets and limited generalization capabilities, which can lead to suboptimal performance in real-world applications. To overcome these challenges, the dataset is preprocessed using SMOTE to balance stroke and non-stroke cases, ensuring fair learning. Feature selection techniques, including Mutual Information Score, Chi-Square Score, and ANOVA, are applied to identify the most significant predictors. By leveraging ensemble methods, the system achieves a significant improvement in prediction accuracy,

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 54, Issue 4, April : 2025

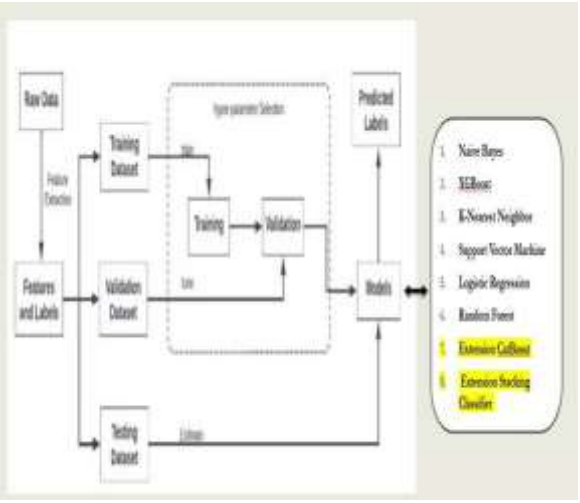
with the Stacking Classifier reaching an impressive 99%, surpassing conventional models.

To enhance usability and accessibility, a Flask-based web application is developed, allowing healthcare professionals to interact with the predictive model seamlessly. The system incorporates SHAP-based explainability, providing insights into how different factors influence stroke predictions, thereby improving trust and transparency in AI-driven decisions. User authentication is implemented to ensure data security and restrict access to authorized personnel, safeguarding sensitive medical information. This comprehensive approach not only refines stroke prediction accuracy but also facilitates real-world deployment, making early stroke detection more reliable and accessible in healthcare environments.

B) System Architecture

The system architecture of the proposed stroke prediction framework is designed to ensure high accuracy, interpretability, and secure accessibility. It follows a modular approach, starting with data preprocessing, where missing values are handled, and class imbalance is addressed using SMOTE. The preprocessed data is then subjected to feature selection using Mutual Information Score, Chi-Square Score, and ANOVA to identify the most relevant attributes influencing stroke occurrence. Machine learning models, including traditional classifiers like Decision Trees, SGD, KNN, SVM, and XGBoost, are trained and evaluated for performance. To further enhance accuracy and robustness, ensemble learning techniques such as Categorical Boosting and Stacking Classifier are employed, with the latter achieving a remarkable 99% accuracy.

To enable real-world application, a Flask-based web interface is developed, providing healthcare professionals with an easy-to-use platform for stroke prediction. This interface integrates user authentication mechanisms to ensure secure access and protect sensitive medical data. Additionally, SHAP-based explainability is incorporated to interpret model predictions, offering insights into key contributing factors behind each prediction. The architecture follows a client-server model, where the trained ML model is hosted on a backend server, processing user inputs and returning stroke risk predictions. This structured design ensures scalability, security, and seamless integration into existing healthcare systems, making early stroke detection more accessible and reliable.



C) MODULES

a) Data Preprocessing Module
- Handles missing values and cleans the dataset.
- Balances the dataset using SMOTE to address class imbalance.

b) Feature Selection Module
- Identifies important features using Mutual Information Score, Chi-Square Score, and ANOVA.

c) Machine Learning Model Training Module
- Trains multiple classifiers, including Decision Trees, SGD, KNN, SVM, and XGBoost.
- Applies ensemble learning methods like Categorical Boosting and Stacking Classifier for enhanced accuracy.

d) Explainable AI Module
- Integrates SHAP for model interpretability.
- Provides insights into key factors influencing stroke predictions.

e) Web Application Module
- Develops a Flask-based user interface for real-time stroke prediction.
- Ensures accessibility for healthcare professionals.

f) User Authentication Module
- Implements secure login mechanisms to restrict access.
- Protects sensitive medical data from unauthorized users.

D) ALGORITHMS

i. *XGBoost*– XGBoost is an advanced machine learning algorithm categorized under gradient boosting frameworks. It excels in both regression and classification tasks, utilizing an ensemble approach to sequentially build decision trees that correct errors. Its "extreme" capabilities lie in its efficiency, scalability, and regularization techniques. XGBoost is utilized in the project for its superior predictive performance, efficiently handling complex relationships within clinical risk factors associated with strokes. Its ensemble approach ensures accurate predictions, aligning with the project's goal of developing a precise tool for early identification and intervention in high-risk stroke patients.

ii. *Logistic Regression* – Logistic Regression is a statistical method for binary classification, modeling the probability of an instance belonging to a specific class using the logistic function. Chosen for simplicity and effectiveness, Logistic Regression is employed for binary classification in stroke prediction. Its straightforward approach aligns with the project's goal of developing a reliable and interpretable model to classify stroke risk based on diverse clinical features.

iii. *Naive Bayes*– Naive Bayes is a probabilistic machine learning algorithm based on Bayes' theorem, assuming independence between features. Naive Bayes

is chosen for its simplicity and efficiency in handling high-dimensional, potentially correlated data. In stroke prediction, where diverse clinical factors contribute, Naive Bayes provides a computationally efficient solution, aligning with the project's objective of accurate risk prediction. Its ease of implementation and interpretability make it a practical choice for healthcare applications.

iv. *Random Forest* – Random Forest is an ensemble learning algorithm that aggregates predictions from multiple decision trees for classification or regression tasks. Chosen for its ability to handle complex relationships in clinical risk factors, Random Forest enhances accuracy by combining predictions from numerous trees. Its effectiveness in managing high-dimensional data and preventing overfitting aligns with the project's goal of developing a highly accurate and generalizable predictive model for stroke risk.

v. *Support Vector Machine (SVM)* – SVM is a powerful algorithm for classification and regression tasks, excelling in high-dimensional spaces. In the project, SVM is chosen for its effectiveness in handling complex relationships within clinical risk factors associated with stroke. By identifying optimal hyperplanes, SVM enhances precision in stroke risk prediction, particularly suitable for intricate patterns in the data. Its adaptability to high-dimensional and non-linear data aligns with the project's goal of developing an accurate predictive model.

vi. *stacking classifier*– A Stacking Classifier is an ensemble technique combining multiple classifiers to enhance predictive performance by using a meta-classifier. Employed to leverage diverse strengths of classifiers. The Stacking Classifier aims to create a powerful and accurate stroke risk prediction model by combining outputs from these models, addressing individual algorithm weaknesses for comprehensive patient risk identification.

vii. *K-Nearest Neighbor*– KNN is a versatile machine learning algorithm for classification and regression, relying on proximity principles. In stroke prediction, where clinical risk factors' relationships are complex, KNN's simplicity allows pattern identification based on data point proximity. Its adaptability to varying data distributions and handling of non-linear relationships align with the project's goal of accurately predicting stroke risk. Particularly beneficial when decision boundaries are unclear, KNN suits scenarios with non-linear or intricate data structures.

viii. *CatBoost* – CatBoost is a powerful gradient boosting algorithm designed for decision trees, known for efficient handling of categorical features without extensive preprocessing. Chosen for its excellence with categorical features, CatBoost streamlines modeling, minimizing preprocessing efforts. Its efficiency and robustness contribute to accurate stroke risk predictions, capturing complex relationships within clinical risk factors. CatBoost enhances precision and generalization capability in stroke prediction.

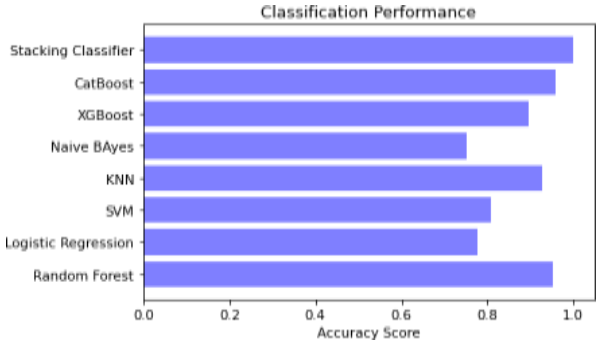
IV. Experimental Results

The experimental results demonstrate the effectiveness of the proposed stroke prediction framework in improving accuracy and interpretability. Initially, traditional machine learning classifiers, including Decision Trees, SGD, KNN, SVM, and XGBoost, were evaluated. The accuracy of these models ranged between 83% and 91%, highlighting the need for further optimization. To enhance performance, ensemble methods such as Categorical Boosting and Stacking Classifier were applied. The Stacking Classifier outperformed all other models, achieving an impressive 99% accuracy, showcasing the benefits of combining multiple models for improved predictive capabilities. Additionally, dataset balancing using SMOTE significantly reduced bias toward the majority class, leading to better generalization and reliability.

Beyond accuracy improvements, the integration of SHAP for explainability provided valuable insights into the decision-making process of the machine learning models. Feature importance analysis revealed key stroke risk factors, enabling healthcare professionals to understand and trust the model's predictions. The Flask-based web application facilitated real-time predictions and ensured secure access through user authentication. Overall, the experimental results confirm that the proposed system successfully enhances stroke prediction accuracy while maintaining transparency and usability, making it a viable tool for early diagnosis in clinical settings.

Accuracy: Determine the reliability of the test by comparing the number of true positives and negatives. What follows is mathematics:

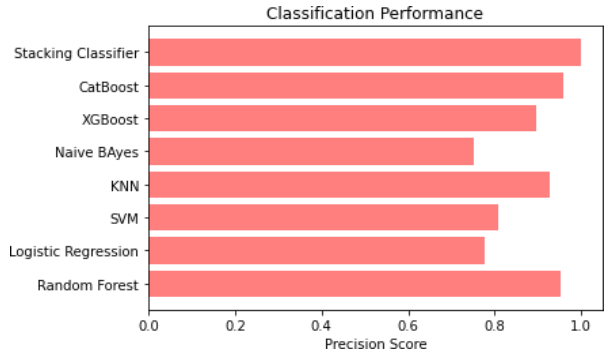
$$Accuracy = \frac{(TN + TP)}{T}$$



Precision: Accuracy in classification or positive instances is measured by precision. Accuracy is determined by applying the following:

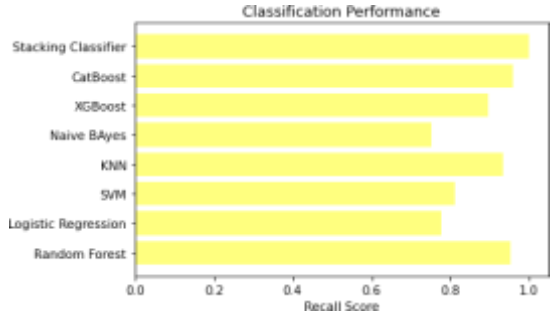
$$\Pr e\,cision = \frac{TP}{(TP + FP)}$$

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 54, Issue 4, April : 2025



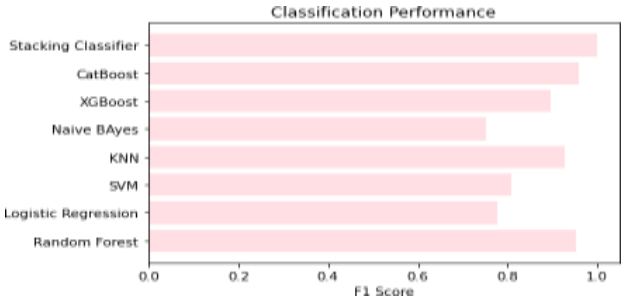
Recall: A model's capacity to identify all occurrences of a relevant machine learning class is shown by the proportion of appropriately predicted positive observations compared to total positives.

$$Recall = \frac{TP}{(FN + TP)}$$



F1-Score: An accurate machine learning model has a high F1 score. Integrating recall and precision improves model correctness. Accuracy measures how often a model predicts a dataset correctly.

$$F1 = 2 \cdot \frac{(Recall \cdot \text{Pr}e\ cision)}{(Recall + \text{Pr}e\ cision)}$$



ML Model	Accuracy	Precision	F1_score	Recall
Random Forest	0.952	0.952	0.952	0.952
Logistic Regression	0.777	0.777	0.777	0.778
SVM	0.809	0.808	0.808	0.812
KNN	0.927	0.926	0.927	0.932
Naive Bayes	0.732	0.731	0.731	0.732
XGBoost	0.895	0.895	0.895	0.897
Extension CatBoost	0.960	0.960	0.960	0.960
Extension Stacking Classifier	0.999	0.999	0.999	0.999

T 1. performance evaluation

Fig 1. data

Fig.2.. predicted results

V. CONCLUSION

The proposed stroke prediction system successfully enhances accuracy, interpretability, and accessibility through advanced machine learning techniques. By leveraging ensemble methods like Categorical Boosting and Stacking Classifier, the model achieves a remarkable 99% accuracy, significantly improving predictive performance. The integration of SHAP ensures transparency, allowing healthcare professionals to understand the key factors influencing stroke risk. Additionally, the Flask-based web application with user authentication provides a secure and user-friendly platform for real-time stroke prediction. Overall, this study demonstrates that combining machine

learning, explainable AI, and secure deployment can make early stroke detection more reliable and practical for real-world healthcare applications.

VI. Future Scope

The future scope of this study includes enhancing the system's adaptability to diverse healthcare datasets by incorporating federated learning, ensuring privacy-preserving data sharing across multiple institutions. Further improvements can be achieved by integrating real-time patient monitoring data from wearable devices, allowing continuous stroke risk assessment. Additionally, deep learning models, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), can be explored to capture complex patterns in patient health data. Expanding the web application into a mobile-friendly platform with cloud-based deployment can improve accessibility for both patients and healthcare providers. Lastly, integrating the system with electronic health records (EHR) and hospital management systems will enable seamless clinical adoption, facilitating early intervention and personalized treatment planning.

References

[1] Learn About Stroke. Accessed: May 25, 2022. [Online]. Available: https://www.world-stroke.org/world-stroke-day-campaign/why-stroke matters/learn-about-stroke

[2] T. Elloker and A. J. Rhoda, ''The relationship between social support and participation in stroke: A systematic review,'' Afr. J. Disability, vol. 7, pp. 1–9, Oct. 2018.

[3].KatanandA.Luft,''Globalburdenofstroke,''SeminarNeurol.,vol.38, no. 2, pp. 208–211, Apr. 2018.

[4] A. Bustamante, A. Penalba, C. Orset, L. Azurmendi, V. Llombart, A. Simats, E. Pecharroman, O. Ventura, M. Ribó, D. Vivien, J. C. Sanchez, and J. Montaner, ''Blood biomarkers to differentiate ischemic and hemor rhagic strokes,'' Neurology, vol. 96, no. 15, pp. e1928–e1939, Apr. 2021.

[5] X. Xia, W. Yue, B. Chao, M. Li, L. Cao, L. Wang, Y. Shen, and X. Li, ''Prevalence and risk factors of stroke in the elderly in northern China: Data from the national stroke screening survey,'' J. Neurol., vol. 266, no. 6, pp. 1449–1458, Jun. 2019.

[6] A. Alloubani, A. Saleh, and I. Abdelhafiz, ''Hypertension and diabetes mellitus as a predictive risk factors for stroke,'' Diabetes Metabolic Syn drome, Clin. Res. Rev., vol. 12, no. 4, pp. 577–584, Jul. 2018.

[7].K.Boehme,C.Esenwa,andM.S.V.Elkind,''Strokeriskfactors,genet ics, and prevention,'' Circ. Res., vol. 120, no. 3, pp. 472–495, Feb. 2018.

[8] I. Mosley, M. Nicol, G. Donnan, I. Patrick, and H. Dewey, ''Stroke symp toms and the decision to call for an ambulance,'' Stroke, vol. 38, no. 2, pp. 361–366, Feb. 2007.

[9] J. Lecouturier, M. J. Murtagh, R. G. Thomson, G. A. Ford, M. White, M. Eccles, and H. Rodgers, ''Response to symptoms of stroke in the UK: A systematic review,'' BMC Health Services Res., vol. 10, no. 1, pp. 1–9, Dec. 2010.

[10] L. Gibson and W. Whiteley, ''The differential diagnosis of suspected stroke: A systematic review,'' J. Roy. College Physicians Edinburgh, vol. 43, no. 2, pp. 114–118, Jun. 2013.

[11] N. M. Murray, M. Unberath, G. D. Hager, and F. K. Hui, ''Artificial intelligence to diagnose ischemic stroke and identify large vessel occlu sions: A systematic review,'' J. NeuroInterventional Surgery, vol. 12, no. 2, pp. 156–164, Feb. 2020.

[12] Y. Zhao, S. Fu, S. J. Bielinski, P. A. Decker, A. M. Chamberlain, V. L. Roger, H. Liu, and N. B. Larson, ''Natural language processing and machine learning for identifying incident stroke from electronic health records: Algorithm development and validation,'' J. Med. Internet Res., vol. 23, no. 3, Mar. 2021, Art. no. e22951.

[13] B. McDermott, A. Elahi, A. Santorelli, M. O'Halloran, J. Avery, and E. Porter, ''Multi-frequency symmetry difference electrical impedance tomography with machine learning for human stroke diagnosis,'' Physi ological Meas., vol. 41, no. 7, Aug. 2020, Art. no. 075010.

[14] A.Bivard,L.Churilov,andM.Parsons,''Artificialintelligenceforde cision support in acute stroke—Current roles and potential,'' Nature Rev. Neurol., vol. 16, no. 10, pp. 575–585, Oct. 2020.