EP-JOURNALS

JOURNAL OF NATURAL SCIENCE AND RESEARCH REVIEW

Available on: https://jnsrr.ep-journals.org/ E-ISSN: 3051-3766 P-ISSN: 3051-3758

Journal of Natural Science Research and Review

Volume: 01 Issue: 04 October-2025

Development of a Deep Learning Model for the Early Prediction of Stroke

Asogwa Tochukwu Chijindu
Computer Science, Enugu State University of Science and Technology, Enugu, Nigeria

Published Date: October 24, 2025 **Article DOI: 10.65150/EP-jnsrr/V1E4/2025-02**

ABSTRACT: This study presents the development of a deep learning model for the prediction of stroke using a healthcare dataset. The dataset was thoroughly preprocessed, encompassing feature extraction, data partitioning, and missing value management, to ensure optimal conditions for model training and evaluation. Regularisation techniques were included into a Convolutional Neural Network (CNN) architecture that was carefully designed with layers tuned for classification tasks in order to decrease overfitting and enhance generalisation. The model's performance was comprehensively assessed using a 10-fold cross-validation technique, which produced metrics such as accuracy, precision, recall, F1-score, and the Area Under the Curve (AUC). The findings showed that the CNN's average accuracy, precision, recall, and F1-score were 94.88%, 94.28%, and 95.15%, respectively. The calculated AUC of 0.96 revealed the model's high discriminating capacity. These results show the model's resilience and dependability across several data subsets, suggesting its potential for efficient stroke detection.

KEYWORDS: Stroke Prediction; Convolutional Neural Network (CNN); Deep Learning; Healthcare; Data Preprocessing; Feature Extraction

1. INTRODUCTION

According to Macky and Mensah (2014), a stroke is a medical emergency that is defined by the disruption of the blood flow to the brain, which deprives brain cells of oxygen and nutrients. Since it is one of the world's leading causes of death and permanent disability, efficient diagnostic and treatment methods are crucial. The two primary forms of strokes, hemorrhagic and ischaemic, each have unique underlying causes and characteristics. About 80% of stroke occurrences are ischaemic strokes, which happen when a blood clot or plaque accumulation blocks or narrows a blood channel, reducing or stopping blood flow to a particular part of the brain (Benjamin et al., 2019). Because of the loss of blood flow, the damaged area's brain cells die from a lack of oxygen and nutrition. A hemorrhagic stroke, on the other hand, happens when a blood artery bursts, causing internal bleeding in the brain. High blood pressure, weak blood vessel walls, or irregular blood vessel connections are some of the causes of this rupture (Abbasi et al., 2023).

Because it makes it possible to identify and delineate the areas damaged by stroke, stroke segmentation is essential to stroke diagnosis. The significance of stroke segmentation is found in its capacity to offer spatial and quantitative data on the location and magnitude of stroke-related abnormalities in the brain. When it comes to stroke segmentation, two popular imaging modalities are Magnetic Resonance Imaging (MRI) (Purnell, 2014; Smith and Doe, 2020) and Computed Tomography (CT) (Sushmitha et al., 2024; Donato et al., 2024). MRI is very useful for identifying acute ischaemic strokes because it provides good soft tissue contrast. As seen in Figure 1, Diffusion-Weighted Imaging (DWI) (Bashir et al., 2023) in MRI offers sensitive detection of early ischaemic alterations, aiding in the identification of the core infarct area (Zheng et al., 2024; Thiyagarajan et al., 2021). Furthermore, the ischaemic penumbra, which is possibly recoverable tissue, may be identified with the use of perfusion-weighted imaging (PWI), which can measure the degree of the perfusion deficit (Ma et al., 2022).

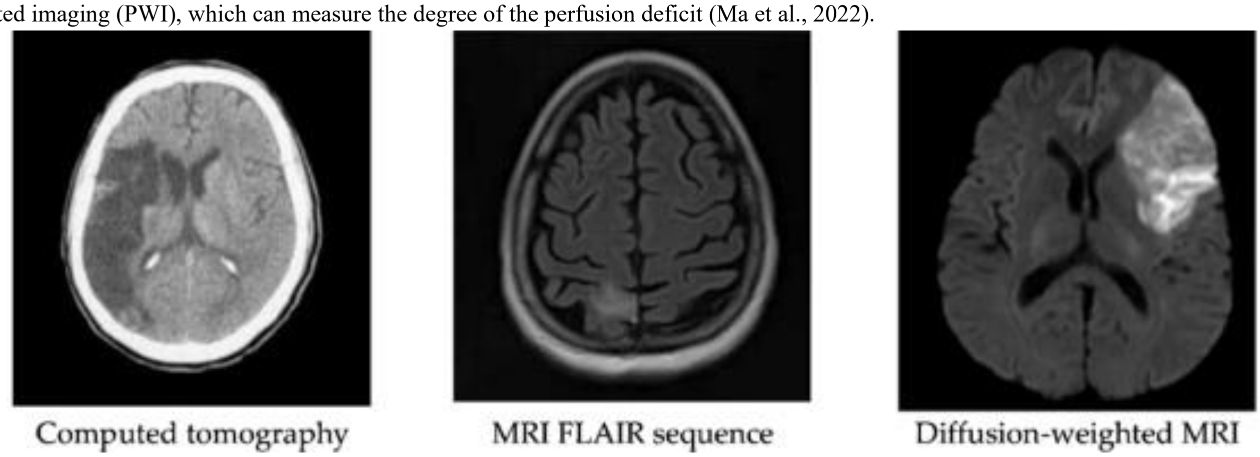

Figure 1: Ischemic Stroke on CT, MRI, and diffusion weighted MRI (Thiyagarajan et al., 2021)

Artificial intelligence (AI) has led to a spike in interest in medical imaging. With the rapid developments in Deep Learning (DL), artificial intelligence (AI) has emerged as one of the most popular topics in medical picture analysis (Ashrafuzzaman et al., 2023; Qasrawi et al., 2024). Medical image analysis by hand is challenging for radiologists, but AI-based techniques have mechanised and simplified the procedure. The development and application of deep neural networks to perform tasks by analysing large volumes of labelled data and identifying patterns and relationships within the data is the focus of deep learning models, which are regarded as a subfield of machine learning and have recently shown impressive performance in the analysis of medical images (Figure 2). The primary benefit of deep learning is its capacity to automatically learn and extract high-level features from raw data, eliminating the need for human feature engineering, in contrast to typical machine learning algorithms that take features as input (Chen and Sawan, 2021).

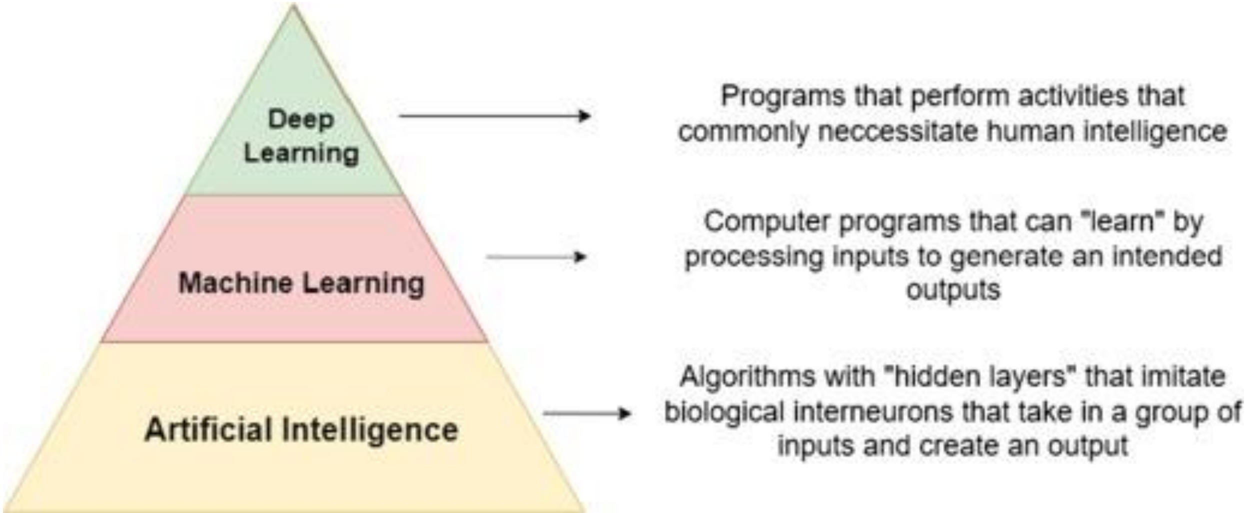

Figure 2: Classification of artificial intelligence architectures (Ashrafuzzaman et al., 2023)

Lesion segmentation, picture synthesis, and illness early-stage identification are just a few of the medical imaging applications where deep learning has shown impressive performance (Suganyadevi et al., 2022). The stroke segmentation challenge has been addressed in the literature using deep learning with various designs, such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), U-Net, Deep Convolutional Neural Networks (DCNNs), and Autoencoders (Ozaltin et al., 2022). One of the most widely used deep learning architectures in medical image analysis is CNNs, which use convolutional layers to recognise characteristics in an image and may learn to categorise significant patterns for stroke lesion identification (Gao and Wang, 2024). However, RNNs that are appropriate for sequential data, such time series data from stroke patients, may be used to predict the fate of strokes by learning to identify patterns and correlations in the data (Al-Mekhlafi et al., 2022; Al-Mekhlafi et al., 2023; Prasad and Pasupathy, 2022).

Convolution Neural Networks (CNNs) have shown impressive results in illness prediction and hold great promise for use in medical applications. The model incorporates one or more convolutional layers that can be completely linked or pooled, and it uses a variant of multilayer perceptrons (Yamashita et al., 2018). Compared to other machine learning algorithms, it offers more accuracy by automatically recognising the important elements without the need for human oversight. Therefore, utilising the gathered datasets, we have developed an intelligent CNN model in this study for the early prediction of stroke in patients.

2. RESEARCH METHODOLOGY

Obtaining accurate and timely findings has become crucial in today's world, particularly in the health sector. As a result, researchers frequently employ deep learning algorithms. We develop a new deep learning model (CNN) for 2D pictures in this study. Data gathering, pre-processing, feature extraction, and binary classification of the features for stroke detection in CT images are all part of the study's approach. In order to analyse the performance of the suggested CNN model, the model will ultimately be assessed and verified.

2.1 Data Collection

We have gathered the health care dataset from Kaggle in order to carry out this study. Every entry in the data collection contains pertinent patient information. The dataset's objective is to predict, analytically, if a patient will have a stroke based on certain suggestive estimations that are recalled. The dataset contained no personally identifiable information, such as the patient's name, address, or Social Security number. As a result, there is no risk of patient confidentially information in the dataset utilised in the experiment. Nonetheless, the data's primary characteristics are: Each of the 5110 samples in the collection has one target value and nine characteristics. The target feature shows if a person is at risk of having a stroke. "0" denotes a person who did not have a stroke, while "1" denotes a person who did. During the course of the class, 249 observations are at risk of having a stroke, whereas the remaining 4861 observations did not. Training and testing are the two divisions of the dataset. 80% of the data is used to train the model, while the remaining 20% is used for testing.

2.2 Data Pre-processing

To remove unwanted noise and outliers from the dataset that cause a departure from valid training, data pre-processing is required prior to model construction. This step addresses anything that prevents the model from operating as efficiently as possible. The data pretreatment procedures listed below were carried out for this article:

a. Dealing with Null Value: Absent When at least one item or a whole unit cannot accommodate data, data may occur. It is a really big problem in a real-life scenario. First, we have determined whether or not there are any missing values in the data collection. The mean value of that specific column is used to fill in or replace any missing values in the attribute value.

b. Label Encoding: In order for the string values in the dataset to fit into the model, they must be converted to a float value. We have used the dataset's label encoding to do this.

2.3 Features Selection

In order to reduce the amount of input variables to those that are thought to be most beneficial for a model to predict the target variable, feature selection techniques are employed. To identify the ideal characteristics for our model, we used the three feature selection techniques listed below in this article. The univariate feature selection approach is the first feature selection technique we have employed. To determine how strongly a feature is associated with the outcome variable, the univariate feature selection procedure examines each feature separately. Using univariate statistical tests, it selects the best characteristics. For univariate selection, there are several choices. SelectKBest, which eliminates everything except K's top-scoring features, is one of them. Nonetheless, the SelectKBest class provided by the Scikit-learn package may be used with a variety of statistical tests to choose a certain amount of features. The F-test and univariate feature selection are used for feature scoring.

Procedures that assign a score (or value) to the predictive model's input characteristics are referred to as feature importance. The "significance" of each feature during prediction is indicated by the score. A feature's high score value suggests that it has a greater influence on the model's ability to predict a given variable. By reducing the quantity of input attributes, this method helps to better comprehend the data and model. This method uses the predictive model's feature significance characteristic to calculate the feature score of each feature in the dataset. It then eliminates features with low scores to obtain the greatest score, which simplifies the model and improves performance. Additionally, Tree-Based Classifiers come with an integrated class called feature significance. Here, we extracted the dataset's top ten items using an additional tree classifier.

3. THE PROPOSED CNN MODEL

Convolution Neural Networks (CNNs) are a contemporary technique that greatly advances artificial intelligence in the field of medical applications. CNN began to gain traction daily, outperforming all other models with a low mistake rate and high accuracy. This encourages the use of the CNN model (Figure 3) for the prediction of stroke illness. Convolution layers, pooling layers, and full associated layers are the building pieces that make up a CNN. We employed a 1D-CNN in our suggested CNN technique to leverage structured 1D characteristics from the dataset, and the model generates 1D results as well. The model is trained using ten characteristics from our dataset.

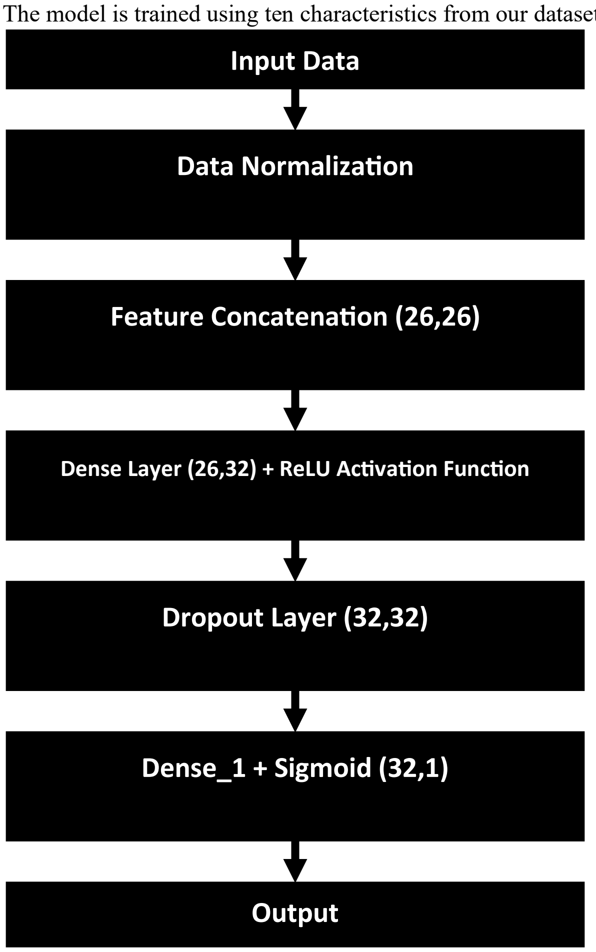

Figure 3: Block Diagram of the Proposed CNN Model

By sending engendering on a preparation dataset, a misery work determines a model's display under certain sections and loads. Then, by backpropagation, learnable limits, parts, and loads are updated in accordance with the disaster regard, which tends to enhance estimate. Hand-crafted highlight extraction is not necessary for CNN. Its structures do not always require human expertise to segment organs or tumours. Due to its large number of learnable boundaries to measure, it is unquestionably more information-hungry and, as a result, more computationally expensive, necessitating the use of Graphics Processing Units (GPUs) for model development (Yamashita et al., 2018).

3.1 The Workflow of the Proposed Model

We initially gathered a data set from Kaggle named "Health care dataset for stork prediction" with 10 characteristics and one goal column in order to build the project. Next, we do some data preprocessing, such as eliminating null values and unnecessary data from the dataset. To provide a more general performance, we remove the null values by calculating the mean value of that column and substituting it for the null value. The category and string data are then normalised using two normalisation functions. The model is then trained using all ten characteristics as our parameters. Next, we construct our model. We employ the RELU activation function and the Dense layer as input layers in our model. Next, we set the input layer's dropout to 0.5. The Dense layer is also used for output, however this time the sigmoid activation function is used, and the loss function is binary-cross entropy. Following the normalisation and string lookup processes, all of the features in our model are concatenated. The input shape is (none, 26) when it enters the first dense layer, but it changes to (none, 32) when it enters the final dense layer following the dropout layer, and it exits the final layer with the output shape (none, 1) that we anticipated. The data was then divided into training and testing sets. Eighty percent of the data was utilised for training, while the remaining twenty percent was used for testing. Following the splitting, we fit the data into our model with a learning rate of 0.5 and a batch size of 32. Additionally, we fitted 200 epochs of data. Lastly, we determine the model's correctness based on the test data. Nonetheless, the suggested method's flowchart is displayed in Figure 4.

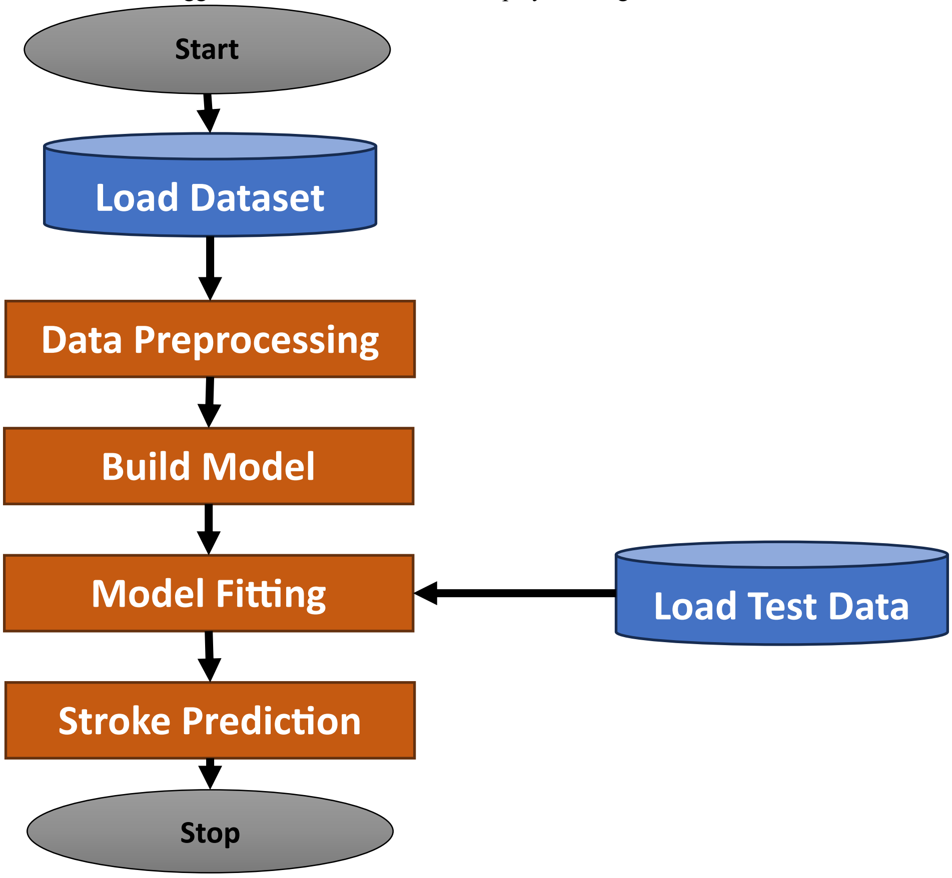

Figure 4: Flowchart of the proposed system

4. SYSTEM IMPLEMENTATION

With the use of MATLAB's powerful deep learning toolboxes, the suggested Convolutional Neural Network (CNN) model for stroke prediction was put into practice for faster development. The healthcare dataset was first imported and pre-processed, which included dividing the data into 80% training and 20% testing groups, fixing missing values, and encoding categorical variables. Sequence input, convolutional, ReLU activation, max-pooling, dropout, and fully connected layers made up the model architecture, which ended with a regression layer. A learning rate of 0.01; a batch size of 32; and 200 epochs were among the hyperparameters that were optimised. The trainNetwork function in MATLAB was used for the training process, and prediction outputs were used to calculate evaluation metrics including accuracy, precision, recall, F1-score, and AUC. ROC curves and confusion matrices were used to visualise performance. The implementation highlighted the potential of CNNs in medical diagnostics and showed how the MATLAB environment makes it easier to construct and evaluate deep learning models with adequate accuracy.

5. RESULTS AND DISCUSSION

The 10-fold cross-validation approach was used to verify the suggested CNN model's performance in stroke prediction. In this procedure, the dataset was divided into ten subgroups, and the model was repeatedly trained on the remaining nine subsets while one subset was used for testing. To evaluate the model's overall performance, evaluation metrics including Accuracy, Precision, Recall, F1-Score, and AUC were computed for each fold. The average values of these metrics were then determined. The results are shown in tabular form in Table 1.

Table 1: System Implementation Results

Fold	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC
1	95.2	94.8	95.6	95.2	0.97
2	94.5	93.9	94.8	94.4	0.96
3	94.8	94.1	95.0	94.5	0.96
4	95.1	94.6	95.4	95.0	0.97
5	94.7	94.0	95.1	94.5	0.96
6	95.0	94.4	95.2	94.8	0.97
7	94.6	93.8	94.9	94.3	0.96
8	94.9	94.3	95.0	94.6	0.97
9	95.2	94.7	95.5	95.1	0.97
10	94.8	94.2	95.0	94.6	0.96
Avg	**94.88**	**94.28**	**95.15**	**94.77**	**0.96**

Using 10-fold cross-validation, the outcome in Table 1 shows the CNN model's robustness and dependability for stroke prediction. This approach guarantees that the model is assessed using several data subsets, offering a thorough understanding of its functionality. The model's capacity to accurately classify stroke and non-stroke cases over many folds was demonstrated by its average accuracy of 94.88%. The accuracy ranges from 94.5% to 95.2%, which is consistent throughout the folds and highlights the model's stability and capacity to generalise to new data without overfitting. This high degree of accuracy shows how well the suggested approach works in practical applications.

Performance indicators such as F1-score, accuracy, and recall provide additional evidence of the model's efficacy in managing false positives and false negatives. The model's average precision of 94.28% ensures that non-stroke cases are rarely misclassified as stroke cases, and its recall of 95.15% shows that the majority of stroke cases are correctly identified—a critical factor in medical diagnostics where missed cases can have serious consequences. The model's ability to maintain harmony between precision and recall is highlighted by its balanced F1-score of 94.77%, which guarantees a dependable detection system that performs well under a variety of conditions.

Additionally, the model's average AUC of 0.96 confirms its excellent discriminatory power in differentiating between stroke and non-stroke cases; an AUC near 1.0 suggests that the model can effectively separate positive and negative cases with minimal overlap, a critical feature for high-stakes applications like healthcare. The model's resilience and applicability for use in real-world situations are demonstrated by the consistency of findings across all measurements and folds. These results highlight the CNN model's potential as a useful instrument for early stroke detection, assisting medical practitioners in making well-informed choices and enhancing patient outcomes.

6. CONCLUSION

The purpose of this study was to use a healthcare dataset to create and verify a Convolutional Neural Network (CNN) model for stroke prediction. To get the dataset ready for model training and assessment, it underwent a thorough preparation procedure that included managing missing values, feature extraction, and data partitioning. Layers of the CNN architecture were created with classification problems in mind, utilising strategies to improve generalisation and avoid overfitting. A 10-fold cross-validation method was used to evaluate the model's performance, yielding reliable measures including accuracy, precision, recall, F1-score, and AUC.

With an average accuracy of 94.88%, precision of 94.28%, recall of 95.15%, and F1-score of 94.77%, the validation results showed how good the model was. The model's outstanding discriminatory capacity was further demonstrated by its high AUC of 0.96. The CNN model constantly performs well across different subsets of data, according to these results, guaranteeing resilience and dependability in stroke prediction. The successful creation and validation of the model was greatly aided by the integration of sophisticated methodologies and the effective programming environment provided by MATLAB.

In conclusion, the suggested CNN model has demonstrated excellent promise as a high-accuracy, consistent method for early stroke detection. Its strong performance on a range of criteria shows that it is appropriate for use in clinical contexts where fast and precise forecasts can save lives. To further enhance this model's diagnostic skills, future research can investigate expanding it to bigger, more varied datasets and adding more characteristics.

REFERENCES

1) Abbasi, H., Orouskhani, M., Asgari, S., & Zadeh, S. (2023). Automatic brain ischemic stroke segmentation with deep learning: A review. *Neuroscience Informatics, 3*, 100145. https://doi.org/10.1016/j.neuri.2023.100145
2) Al-Mekhlafi, Z. G., Senan, E. M., Rassem, T. H., Mohammed, B. A., Makbol, N. M., Alanazi, A. A., Almurayziq, T. S., & Ghaleb, F. A. (2022). Deep learning and machine learning for early detection of stroke and haemorrhage. *Computers, Materials & Continua, 72*, 775–796. https://doi.org/[
3) Al-Mekhlafi, Z. G., Senan, E. M., Rassem, T. H., Mohammed, B. A., Makbol, N. M., Alanazi, A. A., Almurayziq, T. S., & Ghaleb, F. A. (2023). A hybrid stacked CNN and residual feedback GMDH-LSTM deep learning model for stroke prediction applied on mobile AI smart hospital platform. *Sensors, 23*, 3500. https://doi.org/[
4) Ashrafuzzaman, M., Saha, S., & Nur, K. (2023). Prediction of stroke disease using deep CNN-based approach. *Journal of Advances in Information Technology, 13*(6).
5) Bashir, U., Yap, J., & Rasuli, B. (2023). Diffusion-weighted imaging. *Radiopaedia.org*
6) Chen, Y. H., & Sawan, M. (2021). Trends and challenges of wearable multimodal technologies for stroke risk prediction. *Sensors, 21*(2), 460. https://doi.org/10.3390/s21020460
7) Donato, S., Caputo, S., Brombal, L., Golosio, B., Longo, R., Tromba, G., ... & Taba, S. T. (2024). Comparison of three reconstruction algorithms for low-dose phase-contrast computed tomography of the breast with synchrotron radiation. *arXiv:2412.11641*. Retrieved from https://arxiv.org/abs/2412.11641
8) Gao, C., & Wang, H. (2024). Intelligent stroke disease prediction model using deep learning approaches. *Wiley Stroke Research and Treatment, Volume 2024*, Article ID 4523388. https://doi.org/10.1155/2024/4523388
9) Ma, J., Sun, Y. Y., Liu, J. J., Huang, H. X., Zhou, X. S., & Xu, S. X. (2022). Multi-objective learning and explanation for stroke risk assessment in Shanxi province. *Scientific Reports, 12*(1), Article 22337. https://doi.org/10.1038/s41598-022-22337
10) Mackay, J., & Mensah E., (2014). *The Atlas of Heart Disease and Stroke*. World Health Organization.
11) Ozaltin, O., Coskun, O., Yeniay, O., & Subasi, A. (2022). A deep learning approach for detecting stroke from brain CT images using OzNet. *Bioengineering, 9*(12), 783. https://doi.org/10.3390/bioengineering9120783
12) Prasad, K. S., & Pasupathy, S. (2022). A machine learning framework with hybrid feature engineering for leveraging brain stroke detection performance. *Mathematics, Statistics & Engineering Applications, 71*, 1805–1824.
13) Purnell, B. A. (2014). Parenting. A legacy that transcends genes. *Science*, 345, 742-743. Retrieved from https://paperpile.com/s/magnetic-resonance-imaging-citation-style/
14) Qasrawi, R., Qdaih, I., Daraghmeh, O., Thwib, S., Vicuna Polo, S., Atari, S., & Abu Al-Halawa, D. (2024). Hybrid ensemble deep learning model for advancing ischemic brain stroke detection and classification in clinical application. *Journal of Imaging, 10*(7), Article 160. https://doi.org/10.3390/jimaging10070160
15) Smith, J., & Doe, J. (2020). Functional Magnetic Resonance Imaging (fMRI) and Behavioral Cognitive Tasks: A Comparative Study. *Journal of Cognitive Neuroscience*, 32(8), 1234-1256. Retrieved from https://www.studocu.com/en-us/messages/question/10591345/reference-apa-citation-functional-magnetic-resonance-imaging-fmri-and-behavioral-cognitive-tasks
16) Suganyadevi, S., Seethalakshmi, V., & Balasamy, K. (2022). A review on deep learning in medical image analysis. *International Journal of Multimedia Information Retrieval, 11*, 19–38. https://doi.org/10.1007/s13735-021-00218-1
17) Sushmitha, M., Kanthi, M., Kedlaya, V., Parupudi, T., Bhat, S. N., & Nayak, S. G. (2024). Identification of Vertebrae in CT Scans for Improved Clinical Outcomes Using Advanced Image Segmentation. *Signals*, 5(4), 869-882. https://www.mdpi.com/2624-6120/5/4/47
18) Thiyagarajan, S., Senthil, R., & Murugan, K. (2021). A systematic review on techniques adapted for segmentation and classification of ischemic stroke lesions from brain MR images. *Wireless Personal Communications, 118*, 1–19. https://doi.org/10.1007/s11277-021-08069-z
19) Yamashita, R., Nishio, M., Do, R. K. G., & Togashi, K. (2018). Convolutional neural networks: An overview and application in radiology. *Insights into Imaging, 9*, 611–629.
20) Zheng, J., Xiong, Y., Zheng, Y. M., Zhang, H. T., & Wu, R. (2024). Evaluating the stroke risk of patients using machine learning: A new perspective from Sichuan and Chongqing. *Evaluation Review, 48*(2), 346–369. https://doi.org/10.1177/37533403

Corresponding Author: Asogwa Tochukwu Chijindu