An Optimistic K-Nearest Neighbor Algorithm for Detecting Brain Stroke in Comparison with Logistic Regression Algorithm to Improve the Accuracy

A. Ravi Chandra* and Jaisharma K.†
Department of Computer Science and Engineering, Saveetha School of Engineering,
Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamil Nadu, 602105, India

Keywords: Brain Stroke, Health, Death Rate, Machine Learning, Novel Optimistic K-Nearest Neighbor Algorithm, Logistic Regression.

Abstract: The study aimed to improve stroke prediction accuracy using the Novel Optimistic K-Nearest Neighbour Algorithm (NOKNN) and contrast its efficacy with Logistic Regression (LR). The research utilised both the NOKNN and LR algorithms to diagnose brain strokes through machine learning. A sample of 40 was employed for evaluation, split equally between the two algorithms. The aim was to amplify the overall precision of the research. The ClinCalc tool was used for accuracy assessment, set with specific parameters like a 0.8 alpha, 0.8 G-Power, 0.05 significance value, and a 95% Confidence Interval. Results demonstrated NOKNN's superior performance with an accuracy of 98.5%, compared to LR's 93.05%. Following an independent T-test, statistical significance was evident. Advances in technology underline the importance of sophisticated supervised solutions, with NOKNN outperforming LR in this context.

1 INTRODUCTION

The World Health Organisation defines stroke as a risk factor that may increase the death rate or present symptoms persisting for 24 hours or more. A brain stroke is a medical emergency necessitating immediate attention to reduce brain damage and prevent complications. The author of an article posits that ACSL4 plays a pivotal role in the formation of ischemic brain injury, suggesting that regulating its expression could offer a potential therapeutic strategy for stroke management (Cui et al. 2021). In another article (Wu et al. 2018), malignant brain oedema is identified as a grave complication of ischemic stroke, with the aim being to find the most reliable markers and models predicting such oedema post-ischemic stroke. As discussed in a review, research on stroke detection offers invaluable insights, potentially helping identify phytochemicals with clinical efficacy for stroke treatment (Sirsat, Fermé, and Câmara 2020). Another article outlines brain ultrasonography, delving into its principles, techniques, and clinical applications (Robba et al. 2019). For this research experiment, relevant papers from 2018 to 2022 were sourced, primarily from the IEEE Xplore and ScienceDirect databases. IEEE Xplore contributed 249 papers, while ScienceDirect added 176, with a central focus on brain stroke prediction (Vickram, A. S et al 2016). Both IEEE Xplore and ScienceDirect are esteemed as leading databases for such research. Predicting brain strokes is intricate, often involving machine learning and image processing techniques. One innovative approach suggested uses a polymer-based, head-wearable band to collect electromagnetic signals (Alqadami et al. 2019). Another proposal centres on a novel microwave imaging device tailored for brain stroke monitoring, designed for simplicity and user-friendliness (G. Ramkumar et al. 2021; Scapaticci et al. 2018). Further research offers an exploration of phytochemicals with potential therapeutic applications in ischemic stroke, considering their actions and potential benefits (Xu et al. 2021). Identifying specific microRNAs (miRNAs) in the blood to predict stroke risk was another avenue pursued. MiRNAs play a role in disease onset,

* Research Scholar
† Project Guide, Corresponding Author

Chandra, A. and K., J.
An Optimistic K-Nearest Neighbor Algorithm for Detecting Brain Stroke in Comparison with Logistic Regression Algorithm to Improve the Accuracy.
DOI: 10.5220/0012543400003739

In *Proceedings of the 1st International Conference on Artificial Intelligence for Internet of Things: Accelerating Innovation in Industry and Consumer Electronics (AI4IoT 2023)*, pages 37-43
ISBN: 978-989-758-661-3

including strokes (Sonoda et al. 2019) (Deena, S et al. 2022). Of all, the article by (Alqadami et al. 2019) stands out for its groundbreaking system for brain stroke prediction, incorporating wearable electromagnetic head imaging with polymer material. A notable drawback of earlier methods like LR is their limited accuracy and lack of early stroke symptom detection. Conversely, our proposed system seeks to elevate accuracy and facilitate premature stroke symptom prediction, harnessing contemporary algorithms and methodologies. With more accurate data provision, our innovation could catalyse earlier stroke diagnosis, enhancing patient prognosis.

2 MATERIALS AND METHODS

The research was carried out at the Computer Resource Centre Lab of Saveetha School of Engineering, part of the Saveetha Institute of Medical and Technical Sciences. The laboratory boasted high-specification systems utilized for the experiment. The study was divided into two groups: the first group employed the proposed NOKNN algorithm, and the second group used Logistic Regression (LR). The study's total sample size comprised 40 individuals, with 20 samples in each group. Calculations were informed by G-power, adopting an alpha of 0.05 and a beta of 0.95, yielding a 95% CI. These calculations were conducted via the Clincalc website (Stroke Analysis, 2018). For this work's implementation, a Jupyter notebook was used. All codes were executed on a Jupyter notebook on my MacBook Air, which is equipped with 8GB RAM, 256GB SSD storage, and an M1 processor.

2.1 Novel Optimistic K-Nearest Neighbor

The Novel Optimistic KNN Algorithm is a prevalent machine learning technique that employs a distance-based method to discern similarities or differences between two groups within a dataset. For sample classification, the KNN approach often pairs with either the Euclidean or Manhattan distance metrics, thereby classifying new samples into specific groups. Using a dataset with records of stroke and non-stroke patients, the K-NN algorithm identifies the K nearest vectors (or neighbors) to a new sample, subsequently categorizing the new sample into the class (either stroke or non-stroke) with the most akin vectors. The KNN classifier, owing to its simplicity, is a favored algorithm for classification endeavors. At its heart, it calculates the distance between test and training data, subsequently assigning the class of the closest neighbor to the test data. A pivotal parameter in this algorithm is the 'k' value, determining the number of neighbors that will influence the test sample's classification. When k equals one, the test sample's class aligns directly with its nearest neighbor. This methodology is particularly effective with numerical datasets, and various metrics like the Euclidean distance can determine distances between samples. Predominantly, the Novel Optimistic KNN algorithm employs the Euclidean distance.

2.2 Pseudocode

The implementation of the kNN (k-Nearest Neighbors) algorithm in the provided code is delineated as follows:

1. The required libraries are imported: KNeighborsClassifier from sklearn.neighbors for the classification task, from sklearn.metrics for performance evaluation, and pickle for saving and loading the trained model.
2. Empty lists are initialised to record the accuracy, precision, recall, and F1-score values for various k values: these are denoted as acc, ps, rs, and fs respectively.
3. A for loop iterates over k values ranging from 1 to 50.
4. Inside the loop, a KNeighborsClassifier object is instantiated using the current k value and is then trained on the training data (x_train and y_train) via the fit() method.
5. The performance metrics for the given k value are calculated by applying the relevant methods to the test data (x_test and y_test). These metrics are then appended to their corresponding lists (acc, ps, rs, and fs).
6. The trained kNN model is saved to a file labelled "kNN_model" utilising the pickle.dump() method in write-binary ("wb") mode.
7. The index of the highest accuracy within the acc list is determined using the index() method, and this index is stored in the variable, idx. This facilitates identification of the optimal k value.
8. A fresh kNN model is established using the optimal k value, creating a KNeighborsClassifier object with the n_neighbors parameter set to idx+1, and training it on the training data.

9. The class labels for the test data are predicted with the newly trained kNN model, storing the predictions in the variable y_pred.
10. The final trained kNN model is saved to the "kNN_model" file using the pickle.dump() method in write-binary ("wb") mode, overwriting the previously saved model.
11. A classification report, which offers comprehensive details about metrics for every class in the dataset, is presented to gauge the kNN's effectiveness using the classification_report() function.

2.3 Logistic Regression

Logistic Regression (LR) is used as a supervised ML technique to explore the relationship between variables. With LR, it becomes possible to forecast the outcome of the dependent variables. These solutions can be expressed as numerical or binary values, logical bits, and so on. In ML, these values are typically represented in binary format. However, this particular model produces output values that lie between 0 and 1. The key difference between Logistic Regression and Linear Regression is in their applications: Linear Regression addresses regression problems, while Logistic Regression is used for classification tasks using the sigmoid function.

The logistic regression model is expressed as:

$Sigmoid\ Function\ \emptyset(z) = 1 / (1 + e\hat{}-z)$

The testing procedures for the research experiments were undertaken in these steps:

Setting up the environment on Google Colab.

Selecting the runtime type as "py" and choosing the GPU as the hardware accelerator for expedited results.

Importing essential libraries like numpy, pandas, matplotlib, and seaborn.

Loading the dataset link into a variable for data processing.

Conducting operations on the data.

Coding and executing it using the "Run" button.

Monitoring the programme's accuracy by noting the results in an Excel spreadsheet.

Analysing the results in depth with SPSS software.

The dataset used for this experiment, titled Stroke_analysis1, was procured from the Mendeley website, a platform promoting the sharing and access of open-source datasets. The Stroke_analysis1 dataset consists of 14 columns and 4,798 rows, with the columns divided into dependent and independent variables (Stroke Analysis, 2018). These include:

- Patient ID (pid)
- Age
- Gender
- Health Score (nhiss)
- Medical Record Score (mrs)
- Systolic Pressure
- Diastolic Pressure
- Paralysis
- Smoking Status
- Type of Strokes (tos)
- Glucose Levels
- Body Mass Index (bmi)
- Cholesterol Levels
- Risk

2.4 Statistical Analysis

For this study, a dataset was statistically analysed using IBM SPSS v26 software to determine whether there were changes among groups. The dataset included various independent variables: patient ID (pid), age, gender, health score (nhiss), medical record score (mrs), systolic and diastolic blood pressures, paralysis, smoking status, and type of stroke (tos). The study's dependent variables were glucose levels, body mass index (BMI), cholesterol levels, and the risk of stroke. The dataset was sourced from the Mendeley website (Bandi 2020).

3 RESULTS

The difference in accuracy between the Logistic Regression algorithm and the Novel Optimistic K-Nearest Neighbour Algorithm might be due to random chance, suggesting that the Logistic Regression algorithm isn't necessarily more accurate. To ensure an accurate comparison between the two methods, a sample size of 40 was used to test the performance of the algorithms using accuracy measures. Table 1 displays the raw data values for both the proposed and existing algorithms. Each group comprises 20 samples, making a total of 40 samples for both NOKNN and LR.

Table 2 presents the statistics for each group and compares the accuracy of the two different algorithms: NOKNN as the proposed algorithm and Logistic Regression as the existing one. The total number of samples is 40, and the mean accuracy for the proposed algorithm is 98.5136 with a standard deviation of .21446 and a standard error mean of .01469. For the existing algorithm, the mean accuracy

is 93.0636, with a standard deviation of .21446 and a standard error mean of .01469.

Table 1: Outcome of samples when executed. The outcomes of NOKNN and LR are compared to get the average accuracy of each method.

S.NO	NOKNN Algorithm	LR Algorithm
1	98.5	93.05
2	99	93.55
3	98.5	93.05
4	98.3	92.85
5	98.2	92.75
6	98.5	93.05
7	98.6	93.15
8	98.5	93.05
9	98.5	93.05
20	98.5	93.05

Table 3 showcases the independent samples T-Test results for both algorithms. When comparing the accuracy of the algorithms with equal variances assumed, the significance of two-tailed is 1.000, the mean difference is 5.45000, and the standard error difference is .02078. When equal variances aren't assumed, the significance of two-tailed is 0.000, the mean difference remains 5.45000, and the standard error difference is .02078.

Table 2: Group statistics of the data was performed for 20 iterations for the NOKNN and Logistic Regression. The NOKNN (98.5%) outperformed the Logistic Regression.

	Algorithm	N	Mean	Std.Deviation	Std. Error Mean
Accuracy	NOKNN	20	98.5136	.21446	.01469
	LR	20	93.0636	.21446	.01469

Figure 1a presents a bar graph illustrating the average accuracy of NOKNN (98.5136) and the existing algorithm LR (93.0636). The X-axis of the graph represents the algorithms, while the Y-axis displays the accuracy measures achieved. The graph includes 95% confidence intervals and a standard deviation of 1. Figure 1b depicts the architecture diagram for the proposed model, the NOKNN architecture.

Table 3: The independent samples T test of the data was performed for 20 iterations to fix the confidence interval to 95% and statistically significant of p=0.000 ($p < 0.05$).

	Leven's Test for Equality of Variances		T-Test for Equality of Means					95% Confidence Interval of the Difference	
	F	Sig.	t	df	Sig.(2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Accuracy									
Equal Variances assumed	0.00	1.000	262.256	424	0.000	5.45000	0.02078	5.40915	5.49085
Equal Variances not assumed			262.256	424.000	0.000	5.45000	0.02078	5.40915	5.49085

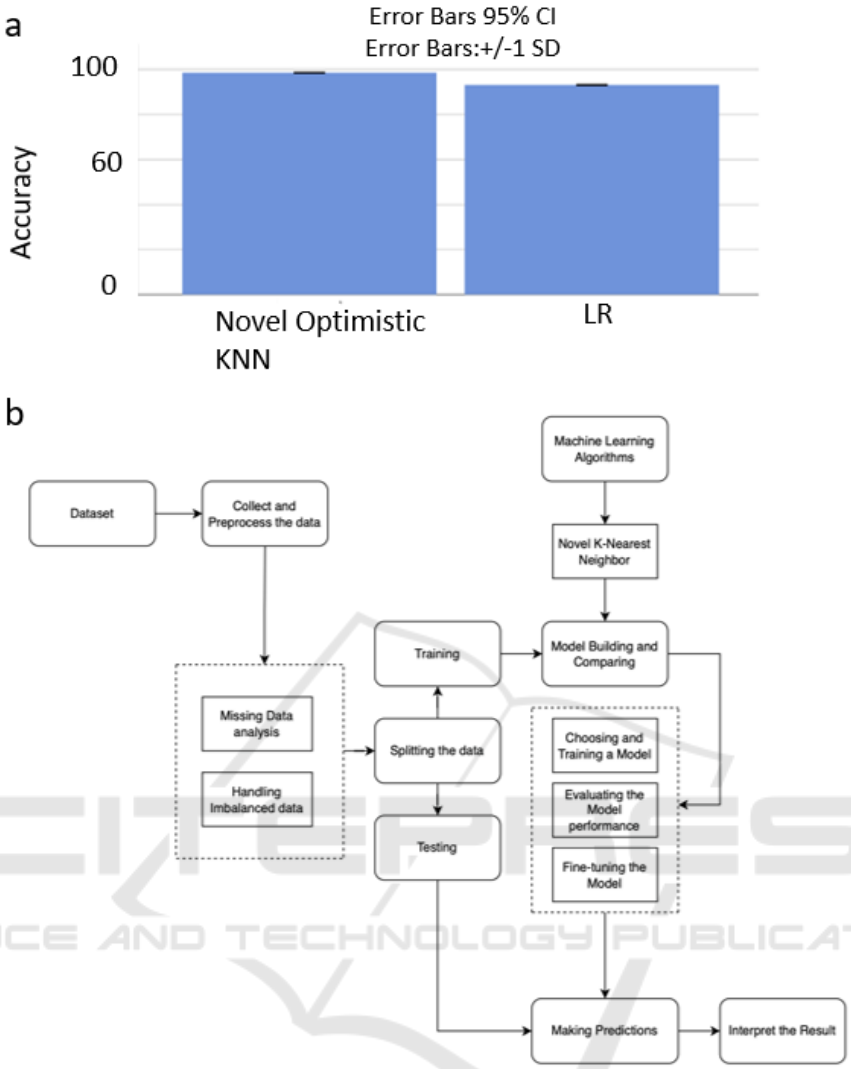

Figure 1: (a)Comparison of NOKNN (98.5%) with Logistic Regression (93.05%) in terms of mean accuracy (b) Classification of unknown samples using a known training dataset by calculating the distance between the two sets.

4 DISCUSSION

The Logistic Regression algorithm achieved an accuracy of 93.0%, which is surpassed by the Novel Optimistic KNN's impressive 98.5%. This notable discrepancy in accuracy suggests that random chance may not be the sole contributing factor. A significance of $p=0.000$ ($p<0.05$) indicates that the performances of the two algorithms are statistically significant, lending robust support to the validity of the test results. The research in this article by Coli et al. (2019) employed microwave tomography, a non-invasive imaging technique that utilizes microwaves to generate images of objects. In another study by Bisio et al. (2018), a technique was introduced to detect and monitor hemorrhagic brain strokes. By imaging the dielectric properties of tissue, it can pinpoint areas of abnormal tissue, such as those affected by a brain stroke and its associated rise in mortality rate. Li et al. (2021) trained a model on a dataset of CT brain images annotated with hemorrhage strokes. Cheng et al. (2020) discussed a Soft Robotic Glove with a Brain-Computer Interface (BCI-SRG), a wearable device utilizing brain-computer interface (BCI) technology to control a soft robotic glove for stroke rehabilitation purposes. Yousif et al. (2021) introduced the Pathological Stroke Classification System (PSCS). The efficacy of the PSCS was assessed using various ML algorithms, including SVM, KNN, and RF. The RF-based

classifier emerged as the most effective, boasting an accuracy of 96.10%. Chourib et al. (2022) highlighted that a Tree-Based Method with Random Forest (TBM-RF) yielded the best results among the feature selection methods examined in their paper. The TBM-RF method achieved an accuracy exceeding 85% and an F1-score over 88% when paired with a decision tree classifier, emphasizing its potency as a feature selection technique, especially when combined with a decision tree classifier.

However, the NOKNN algorithm does come with the drawback of relying on a fixed k value, which can be computationally demanding, potentially resulting in suboptimal runtime performance. Notwithstanding, it offers a superior time complexity when compared to several other machine learning algorithms. Future research suggests a pivot towards enhancing stroke prediction accuracy by integrating image datasets and investigating methodologies across various types of stroke and associated risk levels.

5 CONCLUSION

In recent years, the integration of advanced algorithms in medical diagnostics has transformed patient care, making early detection and timely interventions a reality. This paper's contributions offer insights into the power of modern computational techniques in the arena of stroke prediction. As we move forward, six key points stand out in our exploration:

- Advanced Techniques: The utilization of state-of-the-art techniques and algorithms is crucial in the rapid and precise prediction of stroke, making early interventions possible.
- Superior Performance: The Novel Optimistic K-Nearest Neighbor Algorithm (NOKNN) consistently outperforms traditional methods such as Logistic Regression, showcasing its potential in real-world applications.
- Ensemble Techniques: Combining different methods, as seen with the NOKNN ensemble approach, further refines prediction accuracy, leveraging the strengths of multiple algorithms.
- Risk Factor Identification: Early identification of risk factors can lead to preventative measures, drastically reducing the incidence of stroke in vulnerable populations.
- Comprehensive Analysis: The comprehensive nature of this analysis, which encompasses diverse datasets and conditions, speaks to the versatility of the proposed system.
- Potential for Scalability: With such promising results, there's significant potential to scale and adapt this model across various healthcare settings, offering widespread benefits.

In conclusion, our proposed system marks a pivotal advancement in the realm of stroke prediction. By harnessing the power of sophisticated algorithms, such as NOKNN, we stand at the precipice of a transformative phase in medical diagnostics. Not only does this hold the potential to drastically improve patient outcomes, but it also paves the way for a reduction in the overall death rate. Embracing these advanced techniques, our study underscores the effectiveness of the Novel Optimistic K-Nearest Neighbor Algorithm, achieving a remarkable accuracy of 98.5%. This highlights the necessity of continual research and integration of these tools in the broader medical community.

REFERENCES

Alqadami, Abdulrahman S. M., Konstanty S. Bialkowski, Ahmed Toaha Mobashsher, and Amin M. Abbosh. 2019. “Wearable Electromagnetic Head Imaging System Using Flexible Wideband Antenna Array Based on Polymer Technology for Brain Stroke Diagnosis.” *IEEE Transactions on Biomedical Circuits and Systems* 13 (1): 124–34.

Bandi, Vamsi. 2020. “Stroke_Analysis.” Mendeley Data. https://doi.org/10.17632/jpb5tds9f6.1.

Bisio, Igor, Claudio Estatico, Alessandro Fedeli, Fabio Lavagetto, Matteo Pastorino, Andrea Randazzo, and Andrea Sciarrone. 2018. “Brain Stroke Microwave Imaging by Means of a Newton-Conjugate-Gradient Method in <inline-Formula> L^p </inline-Formula> Banach Spaces.” *IEEE Transactions on Microwave Theory and Techniques* 66 (8): 3668–82.

Cheng, Nicholas, Kok Soon Phua, Hwa Sen Lai, Pui Kit Tam, Ka Yin Tang, Kai Kei Cheng, Raye Chen-Hua Yeow, Kai Keng Ang, Cuntai Guan, and Jeong Hoon Lim. 2020. “Brain-Computer Interface-Based Soft Robotic Glove Rehabilitation for Stroke.” *IEEE Transactions on Bio-Medical Engineering* 67 (12): 3339–51.

Coli, Vanna Lisa, Pierre-Henri Tournier, Victorita Dolean, Ibtissam El Kanfoud, Christian Pichot, Claire Migliaccio, and Laure Blanc-Feraud. 2019. “Detection of Simulated Brain Strokes Using Microwave Tomography.” *IEEE Journal of Electromagnetics, RF and Microwaves in Medicine and Biology* 3 (4): 254–60.

Cui, Yu, Yan Zhang, Xiaolong Zhao, Liming Shao, Guoping Liu, Chengjian Sun, Rui Xu, and Zhaolong Zhang. 2021. “ACSL4 Exacerbates Ischemic Stroke by

Promoting Ferroptosis-Induced Brain Injury and Neuroinflammation." *Brain, Behavior, and Immunity* 93 (March): 312–21.

Deena, S. R., Kumar, G., Vickram, A. S., Singhania, R. R., Dong, C. D., Rohini, K., ... & Ponnusamy, V. K. (2022). Efficiency of various biofilm carriers and microbial interactions with substrate in moving bed-biofilm reactor for environmental wastewater treatment. Bioresource technology, 359, 127421.

G. Ramkumar, R. Thandaiah Prabu, Ngangbam Phalguni Singh, U. Maheswaran, Experimental analysis of brain tumor detection system using Machine learning approach, Materials Today: Proceedings, 2021, ISSN 2214-7853, https://doi.org/10.1016/j.matpr.2021.01.246.

Li, Lu, Meng Wei, Bo Liu, Kunakorn Atchaneeyasakul, Fugen Zhou, Zehao Pan, Shimran A. Kumar, et al. 2021. "Deep Learning for Hemorrhagic Lesion Detection and Segmentation on Brain CT Images." *IEEE Journal of Biomedical and Health Informatics* 25 (5): 1646–59.

Robba, Chiara, Alberto Goffi, Thomas Geeraerts, Danilo Cardim, Gabriele Via, Marek Czosnyka, Soojin Park, et al. 2019. "Brain Ultrasonography: Methodology, Basic and Advanced Principles and Clinical Applications. A Narrative Review." *Intensive Care Medicine* 45 (7): 913–27.

Scapaticci, Rosa, Jorge Tobon, Gennaro Bellizzi, Francesca Vipiana, and Lorenzo Crocco. 2018. "Design and Numerical Characterization of a Low-Complexity Microwave Device for Brain Stroke Monitoring." *IEEE Transactions on Antennas and Propagation*. https://doi.org/10.1109/tap.2018.2871266.

Sirsat, Manisha Sanjay, Eduardo Fermé, and Joana Câmara. 2020. "Machine Learning for Brain Stroke: A Review." *Journal of Stroke and Cerebrovascular Diseases*. https://doi.org/10.1016/j.jstrokecerebrovasdis.2020.105162.

Sonoda, Takumi, Juntaro Matsuzaki, Yusuke Yamamoto, Takashi Sakurai, Yoshiaki Aoki, Satoko Takizawa, Shumpei Niida, and Takahiro Ochiya. 2019. "Serum MicroRNA-Based Risk Prediction for Stroke." *Stroke; a Journal of Cerebral Circulation* 50 (6): 1510–18.

Vickram, A. S., Kamini, A. R., Das, R., Pathy, M. R., Parameswari, R., Archana, K., & Sridharan, T. B. (2016). Validation of artificial neural network models for predicting biochemical markers associated with male infertility. Systems biology in reproductive medicine, 62(4), 258-265.

"Stroke Treatment Prediction Using Features Selection Methods and Machine Learning Classifiers." 2022. *IRBM* 43 (6): 678–86.

Wu, Simiao, Ruozhen Yuan, Yanan Wang, Chenchen Wei, Shihong Zhang, Xiaoyan Yang, Bo Wu, and Ming Liu. 2018. "Early Prediction of Malignant Brain Edema After Ischemic Stroke." *Stroke; a Journal of Cerebral Circulation* 49 (12): 2918–27.

Xu, Hui, Emily Wang, Feng Chen, Jianbo Xiao, and Mingfu Wang. 2021. "Neuroprotective Phytochemicals in Experimental Ischemic Stroke: Mechanisms and Potential Clinical Applications." *Oxidative Medicine and Cellular Longevity* 2021 (April): 6687386.

Yousif, Ahmed Sabeeh, Zaid Omar, Usman Ullah Sheikh, and Saifulnizam Abd Khalid. 2021. "A Novel Pathological Stroke Classification System Using NSST and WLEPCA." In *2020 IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES)*. IEEE. https://doi.org/10.1109/iecbes48179.2021.9398808.