Classification of Ischemic Stroke using Machine Learning Algorithms

Selma Yahiya Adam
Faculty of Mathematical Sciences
University of Khartoum

Adil Yousif
Faculty of Computer Science
University of Science &Technology

Mohammed Bakri Bashir
Faculty of Computer Science &Technology
Shendi University

ABSTRACT
Stroke is the third leading cause of death following diseases of heart and cancer. The majority of strokes are classified as ischemic which have two types: thrombotic and embolic. In thrombotic stroke, the blood clot (thrombus) forms in one of the arteries that supplies blood to brain. The embolic stroke happens when a blood clot that forms somewhere else in the body (embolus) break loose and travels to the brain through the bloodstream. Hemorrhagic stroke is considered another type of brain stroke by some researchers as it happen when an artery in the brain leaks blood or ruptures. As a reason of hemorrhagic stroke, the brain cells damages as result of the pressure from the leaked blood. There are many similarities between these types and it is difficult to classify the cases accurately using medical procedures. Furthermore, there are no clear boundaries between these types. This paper reviewed and analyzed the current studies on classification of ischemic stroke. Furthermore, the study has developed a classification model for ischemic stroke using decision tree algorithm and k nearest neighbor. The classification model is based on a dataset of 400 cases collected from different Sudanese hospitals. The results of the decision tree algorithm can be used by medical specialist to classify and diagnose ischemic stroke patients. Moreover, the study revealed that some features can be used directly to determine the type of ischemic stroke. These results help the medical doctors in the classification process of ischemic strokes. Furthermore, the results found that most of the ischemic stroke cases in Sudan are thrombotic ischemic stroke.

Keywords
Ischemic Stroke, Machine Learning, Decision Tree, KNN

1. INTRODUCTION
Stroke is a blood clot or bleed in the brain which can make permanent damage that has an effect on mobility, cognition, sight or communication. Stroke is considered as medical urgent situation and can cause long-term neurological damage, complications and often death [1, 2]. Stroke is the third leading cause of death following diseases of heart and cancer. The majority of strokes are classified as ischemic which have two types, thrombotic and embolic. In thrombotic stroke, the blood clot (thrombus) forms in one of the arteries that supplies blood to brain. An embolic stroke happens when a blood clot forms away from the patient brain usually in the patient heart and travels through the patient bloodstream to lodge in narrower brain arteries. Hemorrhagic stroke is considered another type of brain stroke as it happen when an artery in the brain leaks blood or ruptures. As a reason of hemorrhagic stroke, the brain cells damages as result of the pressure from the leaked blood. There are many similarities between these types and it is difficult to classify the cases accurately using medical procedures. Furthermore, there are no clear boundaries between these types. This paper reviewed and analyzed the current studies on classification of ischemic stroke. Furthermore, the study has developed a classification model for ischemic stroke using decision tree algorithm and k nearest neighbor. The classification model is based on a dataset of 400 cases collected from different Sudanese hospitals.

One of the main reasons for clot is the fatty deposits that make arteries and lead to a reduced blood flow or other artery conditions. One of the main techniques that is used to diagnose the clot is the brain computed tomography scan, or brain CT scan, which is a test that uses x rays to take clear, detailed pictures of the patient brain[3, 4]. CT scan is mainly done immediately after stroke is suspected. A bleeding in the brain or damage to the brain can be seen using brain CT scan. Other brain conditions that cause patients symptoms can be discovered using brain CT scan. Magnetic Resonance Imaging (MRI) is the second test that is used to examine brain strokes. MRI is based on magnets and radio waves that are used to produce pictures of the organs and structures in the patient's body. Any changes in brain tissue and damage to brain cells from a stroke can be discovered using MRI test. To diagnose a stroke MRI, CT or both can be used[5].

This paper contains seven sections. Section two reviews machine learning algorithms. Section three describes the process of classification of ischemic stroke using machine learning algorithms. The dataset used in the study is described in section four. Section five describes the research methodology. Section six illustrates and discusses the experimentation results. We concluded in section seven.

2. MACHINE LEARNING
The main objective of machine learning methods is to develop computer software that can adapt and learn from their experience. Machine learning is a subfield of artificial intelligence that evolved from the learning process of pattern recognition and computational learning theory[6]. The study is based on two main types of machine learning algorithms decision tree and k nearest neighbor (KNN).

2.1 Decision tree
Decision trees are data structure that has a root node, branches and leaf nodes. Each internal node represents a test on an attribute or feature of the data, each branch represents the outcome of a test, and each leaf node holds a class label. The rood of a tree is located at topmost. "A decision tree is a class discriminator that recursively partitions the training set until each partition consists entirely or dominantly of examples from the same class" [7]. Each branches and root node of the tree contains a split point which is a test on one or more attributes or features and decides how the data is divided and split [7]. The Agrawal and Srikant in [7] described an

example that demonstrate how decision trees works as shown in Figure 1.

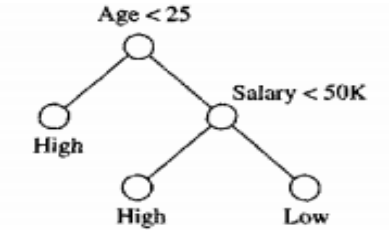

Fig 1: Decision Trees for High and Low credit risk classes Classification[7]

In Figure 1 we can see that (Age < 25) and (Salary < 50K) are two split points that divide the customers fields into high and low credit risk classes. As in this example the decision tree can be employed to predict future applicants by classifying them into the high or low risk classes.

2.2 K nearest neighbor (KNN)

The k-Nearest Neighbors algorithm is a non-parametric technique utilized in classification and regression. When using The k-Nearest Neighbors algorithm for classification and regression, the input consists of the k closest training examples in the feature space[8].

3. CLASSIFICATION OF ISCHEMIC STROKE USING MACHINE LEARNING

The majority of strokes are classified as ischemic which have two types, thrombotic and embolic. In thrombotic stroke, the blood clot (thrombus) forms in one of the arteries that supplies blood to brain. An embolic stroke happens when a blood clot forms away from the patient brain usually in the patient heart and travels through the patient bloodstream to lodge in narrower brain arteries. Hemorrhagic stroke is considered another type of brain stroke as it happen when an artery in the brain leaks blood or ruptures. As a reason of hemorrhagic stroke, the brain cells damages as result of the pressure from the leaked blood[9].

There are many similarities between these types and it is difficult to classify the cases accurately using medical procedures. Furthermore, there are no clear boundaries between these types.

This study employed machine learning algorithms to classify the ischemic strokes. Stroke symptoms that have been used as features for machine learning's classification process are described in the following paragraphs. Brain hemorrhage following an ischemic stroke is a severe difficulty of treatment; yet, its pathology is poorly understood. Using brain imaging for classification may help to better understand and avoid causal factors. Temporary or permanent disabilities may sometimes occur as results of strokes based on how long the brain lacks blood flow and sometimes which part was affected. Other complications such as paralysis or loss of muscle movement may also occur.

The patient may become paralyzed on one side of his body or examine some difficulties in controlling some muscles, for instance those on one side of the patient face or one of his arms. The patient can use physical therapy which can help in returning to activities affected by stroke paralysis, such as walking, eating and dressing. Furthermore, difficulties in talking or swallowing are considered stroke complications as well. A stroke patient may encounter to have less control over the way the muscles in his mouth and throat move, making it difficult for him to talk clearly, swallow or eat. Many stroke symptom and complications have been considered in this research to classify the ischemic stroke accurately as described in the following subsections.

4. THE DATASET

The first step and contribution of this study is generating a data set for ischemic stroke disease. To our best knowledge, the dataset generated by this study is the first dataset for ischemic disease in Sudan. The data set was generated using standard methods for generating benchmarked datasets as described in [10-12]. The dataset items were collected from several hospitals and medical centers in Sudan. The hospital report includes the patient number, age, sex, CT, MRI diagnoses, and other variables for all patients hospitalized in the hospitals participated in the study. The hospitals participated in the study include Alzetona hospital, Antlia hospital and Alneelen center. The data used in the dataset include the data of patient of cases from 2013 to 2015.

The dataset contains 400 patients; their age is mainly between 50 and 88 years. A few cases in the age of 33 years and most of them are male.

Table 1: The features and attributes considered in the study as suggested by specialists in ischemic stroke.

Feature name	The data that the feature contains
A1	The patient number
A2	the age of patient
A3	the sex of patient
A4	if patient have irritability
A5	if patient have convulsions
A6	if patient have left-side weakness
A7	if patient have right-side weakness
A8	patient have mouth deviation
A9	if patient have difficulty in speaking
A10	if patient have unable to walk
A11	if patient have headache
A12	if patient have difficulty in seeing
A13	the result of CT as above
A14	the result of mri as above
A15	the three classes {thrombotic, hemorrhagic embolic}

Table 2: A sample of the dataset

A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11
1	75	female	F	F	T	F	T	T	F	
2	84	male	T	T	F	T	T	T	T	T
3	67	male	F	F	T	F	F	F	T	T
4	55	male	T	F	T	F	F	F	T	T

5	75	male	T	T	F	F	F	F	T	T
6	64	female	T	T	F	F	F	T	T	T
7	60	female	F	F	F	T	F	T	T	T
8	55	female	F	T	F	F	F	F	F	F

A12	A13	A14	A15
F	right lobe acute infarct	right lobe acute infarct	Thrombotic
T	left pontine infarct	left pontine infarct	Thrombotic
T	normal	focal ischemic changes	Thrombotic
T	hypertensive basal ganglia bleed	hypertensive basal ganglia bleed	hemorrhagic
T	hemorrhagic infarct	acute hemorrhagic infarct	hemorrhagic
T	hemorrhagic contusions	hemorrhagic contusions	hemorrhagic
T	ischemic changes	left infarct	Thrombotic
T	ischemic changes	left infarct	Thrombotic

5. RESEARCH METHODOLOGY

This study aims to propose a new model for ischemic stroke classification. The methods employed in this research are split by the six main phases of the research work, as illustrated in Figure 2, which are the problem formulation phase, the dataset collection phase, the preparation of the dataset, the design of the proposed model phase, and the experimentation phase and the results summarizing and discussions.

This research started with formulating the research problem. The formulation of the research problem is performed in two steps: reviewing of the literature and formulating of the research problem. After the research problem formulation, this research identified the scope of the research, the objectives, and limitations of the research procedure. We started this research work with wide and extensive literature review to study the state of the art of the machine learning algorithms as well as ischemic stroke types and classifications.

The second phase of the study is the dataset collection. The data set was generated using standard methods for generating benchmarked datasets as described in [10-12]. The dataset items were collected from several hospitals and medical centers in Sudan. The hospital report includes the patient number, age, sex, CT, MRI diagnoses, and other variables for all patients hospitalized in the hospitals participated in the study.

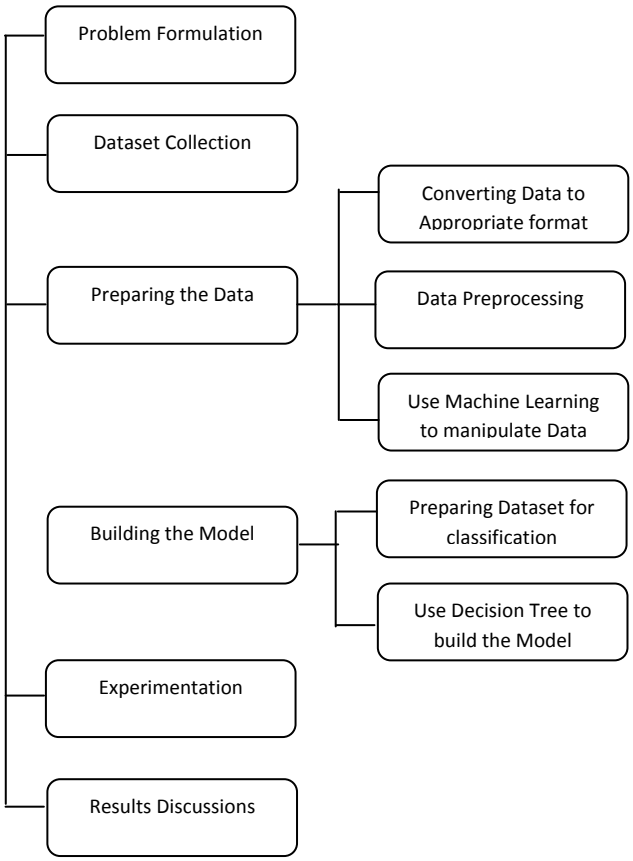

Fig 2: Research Design and Methodology

The third phase of the study was the data preparation which included:

- Converting Data to Appropriate format
- Data Preprocessing
- Use Machine Learning to manipulate Data

Building the proposed model phase contains preparing dataset for classification and using decision tree to build the proposed model.

In the experimentation phase several experiments were conducted and results were collected. The results from the experimentation phase are used in the discussion phase which is the last phase of the study.

6. EXPERIMENTATION RESULTS

This study conducted an experiment using the dataset described in section 3. The study configured the parameters of k- nearest neighbor algorithm as 80% of the dataset for training the data. A classification process for the stroke patients is developed using decision trees and k-nearest neighbor's machine learning algorithms. The developed classification has minimum prediction error as we can see in Table 3. Based on the 14 Stroke attributes as input variables and the three output target stroke values the classification was developed. As shown in Figure 3 the classification is mainly based on attributes of CT scan, MRI test the ischemic stroke types.

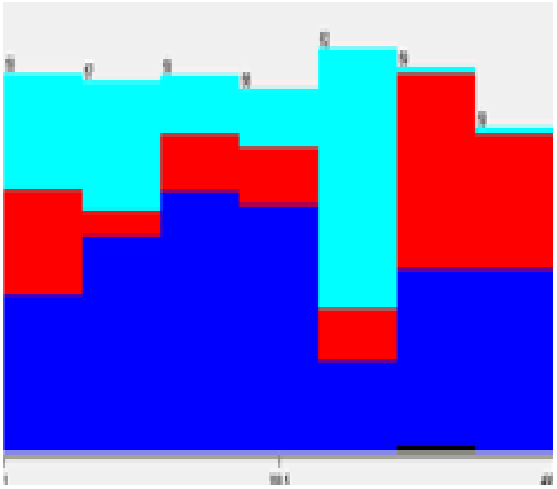

Fig 3: CT scan, MRI test and Ischemic Stroke Types

Figure 4 describes the three classes of ischemic stroke, thrombotic, hemorrhagic and embolic using machine learning algorithm. As we can see in Figure 4 most of the ischemic stroke cases in Sudan are thrombotic ischemic stroke.

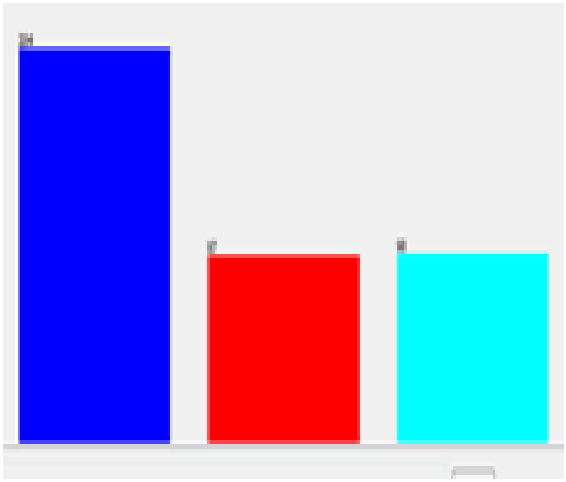

Fig 4: The three Classes of Brain Stroke: Thrombotic, Hemorrhagic and Embolic

Figure 5 and Figure 6 describe the three classes of brain stroke based on CT scan test and MRI test respectively.

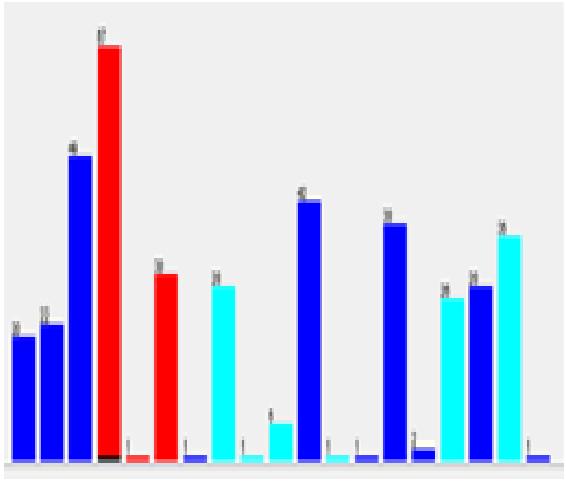

Fig 5: Classification of Brain Stroke using CT Scan

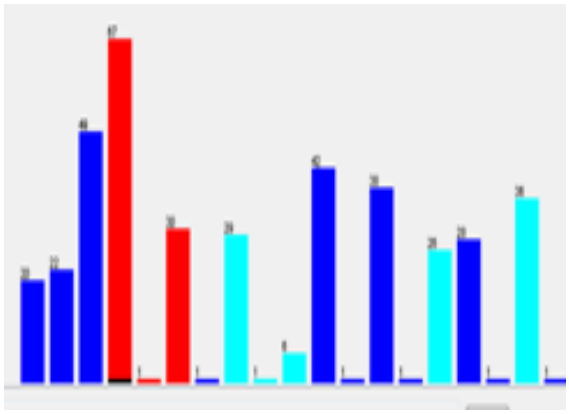

Fig 6: Classification of Brain Stroke using MRI test

Table 4: Detailed Accuracy by Class

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	class
0.975	0	1	0.975	0.987	0.995	Thrombotic
1	0	1	1	1	0.992	hemorrhagic
1	0.018	0.958	1	0.979	0.994	embolic
0.987	0.005	0.988	0.987	0.987	0.994	
0.987	0.005	0.988	0.987	0.987	0.994	Weighted Avg

Table 5: The output using k Nearest Neighbor (knn)

Correctly Classified Instances	Incorrectly Classified Instances	Kappa statistic	Mean absolute error	Root mean squared error	Relative absolute error	Root relative squared error	Coverage of cases (0.95 level)	Mean rel. region size (0.95	Total Number of Instances

								level)	
77	2	0.959	0.0207	0.1294	5.007 %	28.37%	97.468%	33.33%	79
97.468 %	2.5316 %	-	-	-	-	-	-	-	-

Table 6: Detailed Accuracy By Class

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	class
0.975	0.026	0.975	0.975	0.975	0.976	Thrombotic
1	0	1	1	1	0.992	hemorrhagic
0.957	0.018	0.957	1	0.957	0.993	embolic

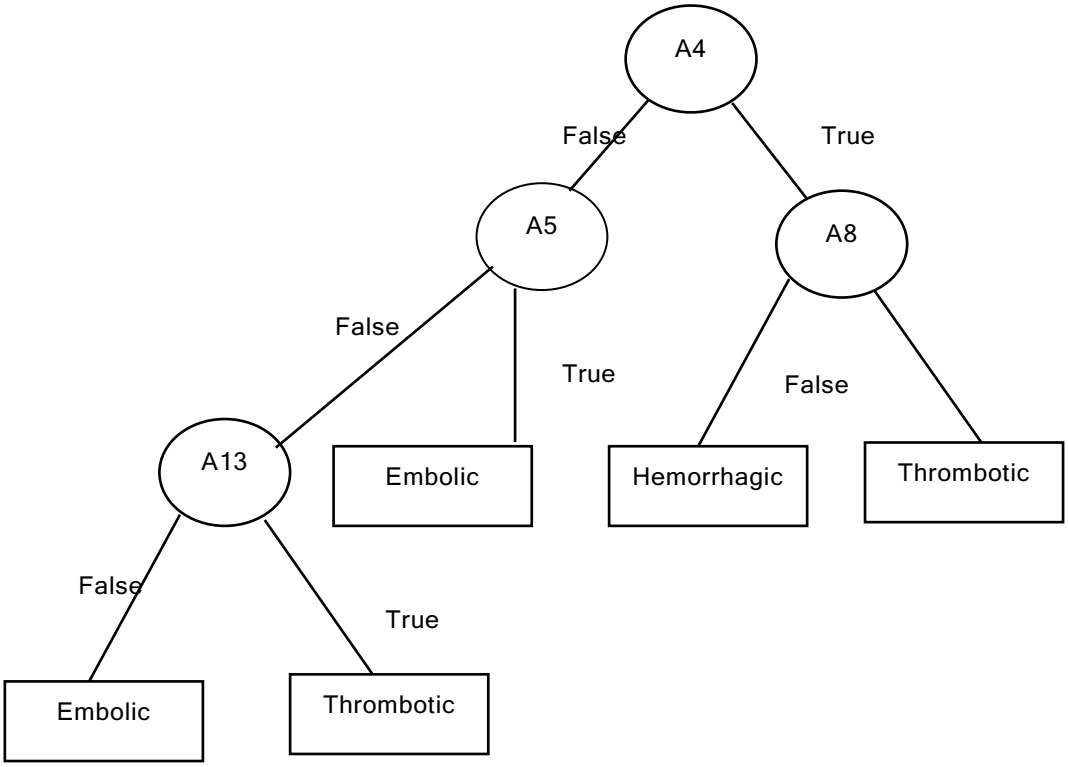

Figure 7: Classification of Brain Stroke using Decision Tree

As shown in Table 3 and subsequent tables the performance of decision tree classification is better than the performance of KNN algorithm.

The results of the decision tree algorithm is a classification model for ischemic stroke that can be used by medical specialist to classify and diagnose ischemic stroke patients. As shown in Figure 7 the features A4(patient have irritability), A5(patient have convulsions), A8(patient have mouth deviation) and A13(the result of CT) can be used directly to determine the type of ischemic stroke. These results help the medical doctors in the classification process.

7. CONCLUSION

There are many similarities between the types of ischemic stroke and it is difficult to classify the cases accurately using medical procedures. Furthermore, there are no clear boundaries between these types. This paper reviewed and analyzed the current studies on classification of ischemic stroke. Furthermore, the study has developed a classification model for ischemic stroke using decision tree algorithm and k nearest neighbor. The classification model is based on a dataset of 400 cases collected from different Sudanese hospitals. The results of the experiment revealed that the performance of decision tree classification is better than the performance of KNN algorithm. The results of the decision tree algorithm can be used by medical specialist to classify and diagnose ischemic stroke patients. The results discovered that the features patient have irritability, patient have convulsions, patient have mouth deviation and the result of CT can be used directly to determine the type of ischemic stroke. These results help the medical doctors in the classification process of ischemic strokes. Furthermore, the results found that most of the ischemic stroke cases in Sudan are thrombotic ischemic stroke.

8. REFERENCES

[1] Ø. Lidegaard, E. Løkkegaard, A. Jensen, C. W. Skovlund, and N. Keiding, "Thrombotic stroke and myocardial infarction with hormonal contraception," New England Journal of Medicine, vol. 366, pp. 2257-2266, 2012.

[2] Ø. Lidegaard, I. Milsom, R. T. Geirsson, and F. E. Skjeldestad, "Hormonal contraception and venous thromboembolism," Acta obstetricia et gynecologica Scandinavica, vol. 91, pp. 769-778, 2012.

[3] D. Gierhake, J. Weber, K. Villringer, M. Ebinger, H. Audebert, and J. Fiebach, "[Mobile CT: technical aspects of prehospital stroke imaging before intravenous

thrombolysis]," RoFo: Fortschritte auf dem Gebiete der Rontgenstrahlen und der Nuklearmedizin, vol. 185, pp. 55-59, 2013.

[4] S. Payabvash, M. H. Qureshi, S. M. Khan, M. Khan, S. Majidi, S. Pawar, and A. I. Qureshi, "Differentiating intraparenchymal hemorrhage from contrast extravasation on post-procedural noncontrast CT scan in acute ischemic stroke patients undergoing endovascular treatment," Neuroradiology, vol. 56, pp. 737-744, 2014.

[5] M. G. Lansberg, M. Straka, S. Kemp, M. Mlynash, L. R. Wechsler, T. G. Jovin, M. J. Wilder, H. L. Lutsep, T. J. Czartoski, and R. A. Bernstein, "MRI profile and response to endovascular reperfusion after stroke (DEFUSE 2): a prospective cohort study," The Lancet Neurology, vol. 11, pp. 860-867, 2012.

[6] J. R. Quinlan, C4. 5: programs for machine learning: Elsevier, 2014.

[7] B. P. Rimal and E. Choi, "A service- oriented taxonomical spectrum, cloudy challenges and opportunities of cloud computing," International Journal of Communication Systems, vol. 25, pp. 796-819, 2012.

[8] M. D. Steenwijk, P. J. Pouwels, M. Daams, J. W. van Dalen, M. W. Caan, E. Richard, F. Barkhof, and H. Vrenken, "Accurate white matter lesion segmentation by k nearest neighbor classification with tissue type priors (kNN-TTPs)," NeuroImage: Clinical, vol. 3, pp. 462-469, 2013.

[9] E. C. Jauch, J. L. Saver, H. P. Adams, A. Bruno, B. M. Demaerschalk, P. Khatri, P. W. McMullan, A. I. Qureshi, K. Rosenfield, and P. A. Scott, "Guidelines for the early management of patients with acute ischemic stroke a guideline for healthcare professionals from the American Heart Association/American Stroke Association," Stroke, vol. 44, pp. 870-947, 2013.

[10] A. Reiss and D. Stricker, "Creating and benchmarking a new dataset for physical activity monitoring," in Proceedings of the 5th International Conference on PErvasive Technologies Related to Assistive Environments, 2012, p. 40.

[11] C. Potter, D. Lew, J. McCaa, S. Cheng, S. Eichelberger, and E. Grimit, "Creating the dataset for the western wind and solar integration study (USA)," Wind Engineering, vol. 32, pp. 325-338, 2008.

[12] A. Chervenak, I. Foster, C. Kesselman, C. Salisbury, and S. Tuecke, "The data grid: Towards an architecture for the distributed management and analysis of large scientific datasets," Journal of network and computer applications, vol. 23, pp. 187-200, 2000.