ISSN: 0975-833X

Available online at http://www.journalcra.com

International Journal of Current Research
Vol. 12, Issue, 04, pp.10948-10952, April, 2020

DOI: https://doi.org/10.24941/ijcr.38446.04.2020

INTERNATIONAL JOURNAL OF CURRENT RESEARCH

RESEARCH ARTICLE

INTEGRATED DATAMINING WITH KNOWLEDGE MANAGEMENT FRAMEWORK FOR STROKE DISEASE

[1]Dr. Anusuya Ramasamy, [2,*]Mr. Gergito Kusse Duba and [3]Mr. Addisu Mulugeta

[1]Assistant Professor, Faculty of Computing and Software Engineering, Arbaminch University, Ethiopia
[2]Lecturer, Department of Computer Science, Debre Tabor University, Ethiopia
[3]Lecturer, Faculty of Computing and Software Engineering,ArbaminchUniversity,Ethiopia

ARTICLE INFO

Article History:
Received 14th January, 2020
Received in revised form
20th February, 2020
Accepted 18th March, 2020
Published online 30th April, 2020

Key Words:
Stroke Disease Diagnos,
JRIP classification, Algorithm,
Data Mining, Stroke,
Knowledge Base System.

ABSTRACT

Stroke disease is a medical condition caused due to inadequate supply of blood to the brain cell that damages the cell and may result in death. In developing country like Ethiopia, the death of stroke patient increases from year to year due to the scarcity of specialists and health facilities. This lack of effort to address such a problem, this research study focuses to design and develop a prototype system by integrating data mining results with knowledge-based system that facilitate diagnosis and treatment for a patient and provides an advice and risk level for the patient. Mixed research design and an integrated knowledge acquisition method were used to acquire knowledge. Orange and WEKA tool were used as hybrid data mining tool to preprocess, analyze datasets and designing the prediction model. About six classification algorithms were comparatively analyzed and finally JRIP classification algorithm has been registered with the better accuracy of 94.16% under 10-fold cross-validation. Rule-based knowledge representation technique was used to represent knowledge in the knowledge base, SWI-Prolog was used to construct knowledge base, Java NetBeans was employed to design GUI for the KBS, JPL library was used as a middleware between knowledge base and designed GUI. Finally, after the system has scored 90% system performance and 89.9% user acceptance which is a promising result that achieves the objective of the study.

Citation: ***Dr. Anusuya Ramasamy, Mr. Gergito Kusse Duba and Mr.Addisu Mulugeta.*** **2020.** "Integerated datamining with knowledge management framework for stroke disease", *International Journal of Current Research*, 12, (04), 10948-10952.

INTRODUCTION

The development of ICT plays a vital role in the day to day activity. Due to the rapidly growing and use of ICT in different business organizations and industry for collecting data related to their own operation. ICT has been generated a large number of databases and huge data in various areas (Rajalakshmi,). The research in databases and information technology has given rise to an approach to store, retrieve and manipulate this important data for further decision making and problem-solving. While the database technologists have been focusing on efficient means of storing, retrieving, deleting and manipulating data, the machine learning community has focused on developing techniques for learning and acquiring knowledge from the huge amount of data. At a time, data can be considered to be a gold mine for strategic planning for research and development in this area known as Data Mining or Knowledge Discovery in Databases (Jain, 2014).

Corresponding author: **Mr. Gergito Kusse Duba,**
Lecturer, Department of Computer Science, Debre Tabor University, Ethiopia.

According to Fayyad (Fayyad, 1996) "Data Mining is a process of finding models, interesting trends or patterns in the large dataset in order to guide the decision about future activities depending on previous and current data". Rajakshmi (Rajakshmi, 2003) defined "Data mining is a process of extraction of useful information and patterns from huge data". Knowledge-Based System is also one part of the study. In this study knowledge-based system is used for the purpose of diagnosing and treatment of the stroke disease patient while the data mining part is used as designing a predicting model of the disease. According to the E.A. Feigenbaum and R.S. Engelmore (Engelmore, 1993) knowledge-based system is a computer program that is a subfield of computer science study known as Artificial Intelligence. The systematic goal of AI is understanding intelligence by developing computer programs that show intelligent behavior about the specific problem.

STATEMENT OF PROBLEM

According to the world health organization international collaborative study (Almadani, 2018), stroke is defined as "a rapidly developing clinical sign of local or global disturbance of cerebral function, lasting more than 24 hours or leading to death with no apparent cause other than that of vascular origin".

S. No.	Research Title and	Authors (Year)	Major Contribution	Critical remark
1	"Integrating Data Mining with Knowledge-Based System for Diagnosis and Treatment of Cattle Diseases: The Case of Debre Berhane Basso Animal Health Center."	Tadesse Beyene (DBU), (Jun 2018 G. C.)	3627 Datasets of cattle disease, Rule-based knowledge representation, J48 classifier Algorithms, SWI-prolog WEKA, JAVA NetBeans and 85% user acceptance testing result	Apply Integration on other domains, 5047 datasets were used, apply hybrid data mining tools, Jrip classifier Algorithm used.
2	"Developing a Knowledge Base System for Diagnosis and Treatment of Malaria"	Chala D., Million M. (PhD), Debela T., (July 2016)	Knowledge Engineering research design, manual knowledge acquisition techniques, Rule-based knowledge representation.	Apply data mining techniques, another domain area, Integrated knowledge acquisition, Java NetBeans for GUI design.
3	"Integrating Data Mining Results with the Knowledge Base System for Diagnosis and Treatment of Visceral Leishmaniasis"	Tesfamariam M., (January 2015 G.C)	2682 dataset used, rule base knowledge representation, system performance result 95% and user acceptance result 86% for a validating system, comparatively analyzing three classifiers with PART classifier 67.5% better result.	5047 datasets in other domain, 90% and 89.8% for system performance and user acceptance respectively for prototype validation, comparatively analysis six classifier and finally Jrip as a better result with 94.16%
4	"Towards Integrating Data Mining with Knowledge-Based System: The Case of Network Intrusion detection."	Abdulkrem M., (June 2013)	614,447 network datasets then applying data reduction, system performance (80%) and user acceptance (88.6%), recommend designing GUI for feature study.	5047 datasets in other domains without data reduction, system performance (90%) and user acceptance (88.6%), design GUI for the prototype system.
5	"A self learning knowledge-based system for diagnosis and treatment of diabetes."	Solomon G., (January 2013)	Rule-based knowledge representation, Manual knowledge acquisition (interview and document analysis), 83.33% system performance and 85% user acceptance.	Apply integration of UMI with KBS on other domains, used integrated knowledge acquisition, 90% system performance and 89.8% user acceptance. designed GUI for the system.

Table 1. Related works

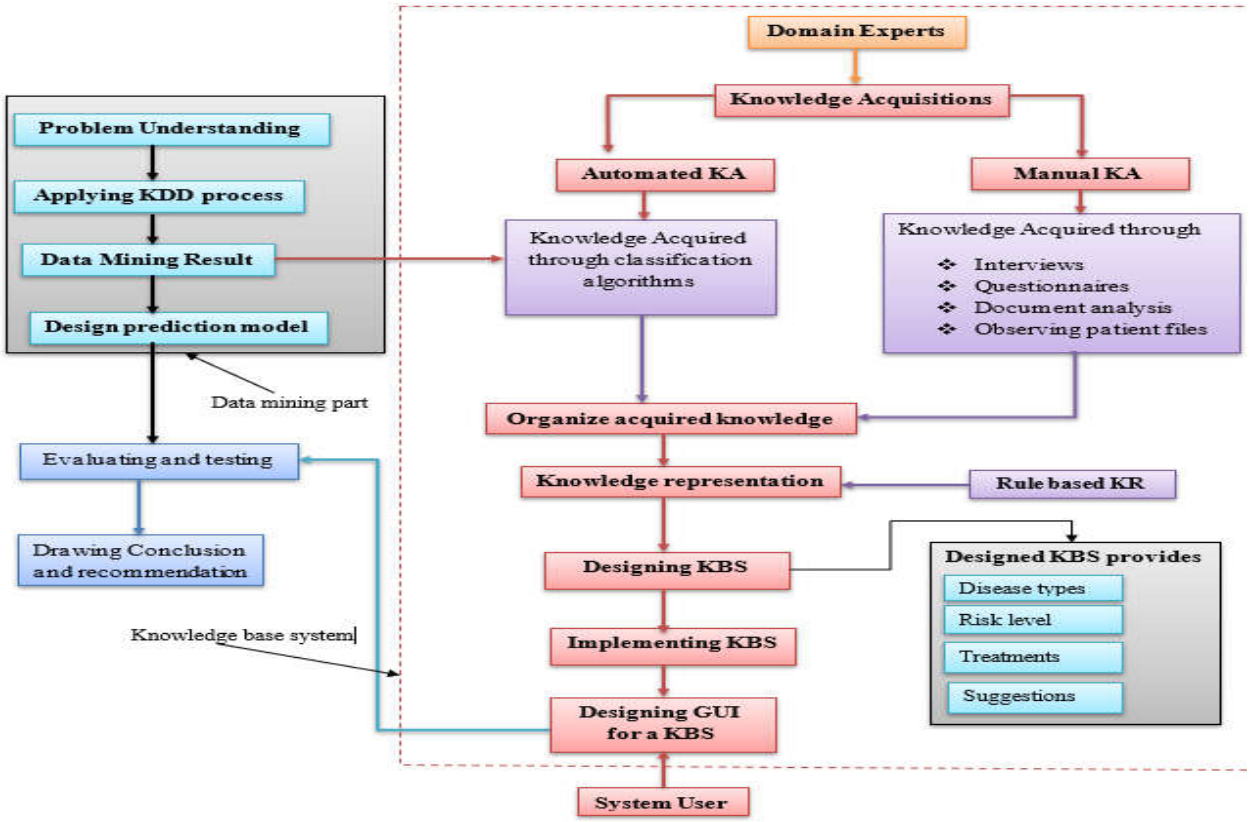

Figure 1. Integrated Datamining and Knowledge Management framework

Classifiers Name	Test Option	Correctly classified		Incorrectly classified		Time in second
		Number	In %	Number	In %	
Bayes Net	Cross Validation	4054	80.34%	993	19.66%	0.02
	Percentage Splitting	1243	83.9%	238	16.1%	0.01
Naïve Bayes	Cross Validation	4013	79.5%	1034	20.5%	0
	Percentage Splitting	1220	82.4%	260	17.6%	0
JRip	**Cross Validation**	**4743**	**94.16%**	**303**	**6.0%**	**1.5**
	Percentage Splitting	**1375**	**93.2%**	**106**	**7.15%**	**1.09**
Decision Table	Cross Validation	4507	89.3%	540	10.7%	0.45
	Percentage Splitting	1329	89.8%	151	10.2%	0.39
J48	Cross Validation	4752	94.0%	295	5.84%	0.12
	Percentage Splitting	1386	92.85 %	95	6.4%	0.05
Random Forest	Cross Validation	4700	93.1%	347	6.9%	1.17
	Percentage Splitting	1354	91.4%	127	8.6%	0.68

Table 2. Performance of the JRip classifier

Classifier	Test Option	Class	Detail Accuracy					
			TP	FP	Perc.	Rec	FM	ROC Area
Bayes Net	Cross Val.	Normal (0)	0.62	0.05	0.86	0.622	0.72	0.94
		Hemorrhagic (1)	0.87	0.12	0.79	0.87	0.83	0.95
		Ischemic (2)	0.92	0.13	0.77	0.92	0.85	0.96
	Percentage Splitting	Normal (0)	0.654	0.040	0.886	0.654	0.753	0.953
		Hemorrhagic (1)	0.891	0.106	0.807	0.891	0.847	0.959
		Ischemic (2)	0.922	0.117	0.807	0.922	0.861	0.960
JRip	Cross Val.	Normal (0)	0.950	0.047	0.910	0.950	0.930	0.959
		Hemorrhagic (1)	0.942	0.027	0.947	0.942	0.994	0.982
		Ischemic (2)	0.927	0.017	0.965	0.927	0.946	0.976
	Percentage Splitting	Normal (0)	0.932	0.057	0.885	0.932	0.908	0.936
		Hemorrhagic (1)	0.947	0.033	0.934	0.947	0.941	0.981
		Ischemic (2)	0.913	0.013	0.973	0.913	0.942	0.968
J48	Cross Val.	Normal (0)	0.912	0.022	0.955	0.912	0.933	0.952
		Hemorrhagic (1)	0.965	0.032	0.937	0.965	0.951	0.986
		Ischemic (2)	0.948	0.034	0.934	0.948	0.941	0.982
	Percentage Splitting	Normal (0)	0.932	0.022	0.954	0.932	0.943	0.963
		Hemorrhagic (1)	0.952	0.032	0.936	0.952	0.944	0.916
		Ischemic (2)	0.945	0.028	0.945	0.945	0.945	0.987

Table 3. Evaluation results of JRIP Classification Algorithm

Worldwide there are about 4.6 million deaths from stroke each year (Bontia, 1985), and this makes it the leading cause of death and disability. Nowadays diagnosis and prediction of stroke disease are totally done manually by the radiologist. In health care centers a large amount of data becomes available. On such a large amount of data, it is very difficult for a human being to manually process these data in a short period of time for the effective patient diagnosis, and treatment schedule. Hence, they are also having less accuracy during the time of treating and diagnosing the patient and also requires an intensively trained person to avoid diagnosis error. As a solution, the proposed study will present a new approach to design a framework for integrating the data mining prediction model with a knowledge base system that supports the diagnosis and treatment of stroke diseases. This study reviewed different researcher research papers in the world on the proposed study.The summary of reviewed papers are summarized in following Table 1. As it has been observed from the strong analysis of various research findings, there is critical research gap. This forces to conduct the study to design a framework to integrate data mining results with a knowledge base system that supports for the stroke disease diagnosis and treatment.

RESEARCH DESIGN AND METHODOLOGY

Research design: The study selected hybrid research design (hybrid form of qualitative and quantitative research design). because during research activity of this study in same part researcher uses quantitative like data collection and sampling techniques research activities and during time of experimental discussion qualitative research design is suitable.

Data Collection: This research study used both primary and secondary data as source of information. The primary stroke patient data are collected from various hospitals specially in south region of Ethiopian (Arba Minch Hospital, Jinka hospital, Hawassa Referal Hospital, and Sodo Christian Hospital). Secondary data are collected through questionnaires and direct observation with checklist. Additionally, different published, public or private documents, books, journals articles, research findings, reports, manuals published by different organizations like WHO and online materials were used as a secondary data source.As addressed in the previous section data were collected as primary and secondary datasets. 5047 datasets were collected from four different hospitals. The following diagram show details of collected datasets with their class type of the diseases,For conducting the proposed research study researcher uses various tools for different purposes like for documentation, data analysis, design models, implementation, and coding. By using comparative analysis of data mining tools orange and WEKA tool was used for preprocessing and designing prediction models, E-draw Max for design models and other research activities, Java netbeen and SWI-prolog for implementing as well as designing the KBS, and JPL(java prolog library) for integrating GUI designed through Java with Prolog programming

Data preprocessing: To get better accuracy data mining results researcher preprocesses the collected data sets. For data preprocessing step researcher uses an orange software tool. In an orange tool to preprocess the data, it is simply done by drag and drop the widget of the tool. The following figure shows the preprocessing of a dataset using an orange It shows that about

0.6% (303) of data sets are missing values and the next figure shows after filling the missing values.

Feature Selection: Medical data is often very high dimensional. Depending upon the use, some data dimensions might be more relevant than others.
In processing medical data, choosing the optimal subset of features is such an important issue in a research study, not only to reduce the processing cost but also to improve the usefulness of the model built from the selected data.

Class unbalancing: The class imbalance problem is occurring in many applications. In the selected dataset there are three class with Normal 502(10%), Hemorrhagic 1817 (36%), and Ischemic 2728 (54%) which imbalanced classes. Class Imbalance is typically occurring when, in a classification problem, there are many more instances of some classes than others. In such cases, standard classifiers tend to be overwhelmed by the large classes and ignore the small ones. Particularly, they tend to produce high predictive accuracy over the majority class, but poor predictive accuracy over the minority class Invalid source specified.. The following figure show number of class before class balancing.

Data mining classifier: To select better accurate algorithms for a prediction model, researcher selects three classifiers and two algorithms from each classifier then uses comparatively analysis the algorithms.

Integration of Data mining result with Knowledge Base System: The aim of this study is to integrate data mining result for the development of a knowledge base system. It is obvious that a knowledge base is a core for a certain knowledge base system. For that knowledge, the acquisition is done using the JRip rule induction algorithm, which performs best for the given stroke disease dataset. The following diagram show the conceptual design for the integration process. To design and implement the knowledge base system researcher uses SWI-prolog and Java NetBeans to design user interface and JPL library for the integration. The following figure show the system architecture for integrating data mining result with knowledge base system. Researcher try to design the prototype to show how the integrated system will work whenever implemented. The system prototype has the knowledge base which provides explanation for the physicians, if he/she didn't understand the question and finally identifies disease type, treatment and risk level of a patient. The user interface facilitates the communication between the knowledge base system and the user. When KBS of stroke disease diagnosis and treatment identified a type of stroke disease, the system would present risk levels, treatments, and preventions in figure 1.

EXPERIMENTAL RESULTS AND DISCUSSION

Experimental setup: A total of two experiments for each algorithm aiming at building predictive models are undertaken. The sampled data set contains 5047 instances containing normal, Ischemic, and Hemorrhagic stroke. The data set contains 11 attributes and all of them are involved in all experiments. Default value of parameters is taken into consideration for each classifier algorithm since it allows achieving better accuracy compared to modifying the default parameters values. In any data mining research before developing a model, Researcher generate a mechanism to test the model performance. For instance, in the supervised data mining task, such as classification, it is common to use classification accuracy measure (CA), True Positive rate (TP), precision, recall and F-measure of the experts are used as to measure the performance of the developed data mining model.

Creating Classification Models Using Different Classifier Algorithm: Bayes Net This experiment conducts under 10-fold cross-validation test option with default parameters of WEKA and the algorithm generates a model as Bayes Net and Correctly Classified Instances are 4054 which means 80.34 % and Incorrectly Classified Instances are 993 which means 19.66% from Total Number of 5047 of Instances and taking 0.02 seconds to build the model.The Bayes Net learning algorithm with percentage split scored classification accuracy out of 1481 number of testing instances 1241 (83.9 %) of them are classified correctly and the remaining 238 (16.1%) testing instances are misclassified or incorrectly classified with time taken 0.01 second to build model. Naïve Bayes This experiment conducts under 10-fold cross-validation test option with default parameters of WEKA and the algorithm generates a model as Naïve Bayes and Correctly Classified Instances are 4013 which means 79.5% and Incorrectly Classified Instances are 1034 which means 20.5% from Total Number of 5047 of Instances and taking 0 seconds to build the model.The Naïve Bayes learning algorithm with percentage split scored classification accuracy out of 1481 number of testing instances 1220 (82.4%) of them are classified correctly and the remaining 207 (17.6%) testing instances are misclassified or incorrectly classified with time taken 0 second to build model. To conclude, the above two experiments namely experiment I and II performed in order to build the classification model using Bayes Net classification algorithm by applying k-fold cross validation and percentage split method in respectively on the experiments

Performance based Comparisons for Proposed Models: The Decision Table learning algorithm with percentage split scored classification accuracy out of 1481 number of testing instances number of testing instances 1329 (89.8%) of them are classified correctly and the remaining 152 (10.2%) testing instances are misclassified or incorrectly classified and 0.39 second time taken to build model is represented in figure 2. Selecting a better classification technique for building a model, which performs best in handling the classification, is one of the aims of this research. For this reason, the three selected classification model with respective best performance accuracy is listed in below table 2. The results of the algorithms are evaluated based on correctly classified instance and incorrectly classified instances. Prediction accuracy shows us the general classification accuracy of the algorithms. Apart from prediction accuracy, classifiers are also evaluated to measure how they correctly classified each class to their correct class or incorrectly classified to another class. Hence, to evaluate the performance of the classifiers employed in this study True Positive rate, False Positive rate, Precision, Recall, F-measure, and ROC Area are used. Table 2 illustrates the performance of the first three classifiers.

Classification Models Using Rule classifier: JRip This experiment conducts under 10-fold cross-validation test option with default parameters of WEKA and the algorithm generates a model as JRip and Correctly Classified Instances are 4743 which means 93.98 % and Incorrectly Classified Instances are 303 which means 6.02% from Total Number of 5047 of

Instances and taking 1.5 seconds to build the model. The JRip learning algorithm with percentage split scored classification accuracy out of 1481 number of testing instances 1375 (92.8%) of them are classified correctly and the remaining 106 (7.15%) testing instances are misclassified or incorrectly classified and 1.09 second time taken to build model.

Decision Table This experiment conducts under 10-fold cross-validation test option with default parameters of WEKA and the algorithm generates a model as Decision Table and Correctly Classified Instances are 4507 which means 89.3 % and Incorrectly Classified Instances are 540 which means 10.7% from Total Number of 5047 of Instances and taking 0.45 seconds to build the model. Prediction accuracy shows us the general classification accuracy of the algorithms. Apart from prediction accuracy, classifiers are also evaluated to measure how they correctly classified each class to their correct class or incorrectly classified to another class. Hence, to evaluate the performance of the classifiers employed in this study True Positive rate, False Positive rate, Precision, Recall, F-measure, and ROC Area are used. Table 2 illustrates the performance of the first three classifiers. J48 with Percentage split test option has registered the best result in terms of the precision, recall and F-measure values as compared to other classifiers all over the three classes. the following table 3 show detailed accuracy for the first three classifier.

With regard to the classification accuracy listed above the table information from the experiment J48 scores the highest classification accuracy with 94.16% the second highest accuracy values is scored by JRip algorithm with 39 generated rule. JRip and J48 have registered almost similar FP rate values for all three class. Baye Net has registered the Largest FP rate (13%) for Ischemic class with cross validation test option as compared to the other two algorithms. The rule acquired from the classifier algorithms is used for constructing knowledge base. So as to develop effective knowledge base system, acquiring relevant rules is critical issues. Hence from the six algorithms the researcher selected the classifier which best performed on classifying the data set. In table above according to experimental result J48 and JRip algorithms scores the highest accuracy.JRIP classifier has generated 39 rules. The rules involved 10 features/attributes among the 11 features/attributes from the sample data set. The algorithm generated 27 rules for Normal class, 11 rules for Hemorrhagic class, and only one rule for Ischemic. In consultation with domain experts in the area of stroke specialized, the rules are evaluated to make sure that whether or not they tell us about stroke behaviors. Based on the evaluation, the rules are capable of identifying stroke type but question is raised that the algorithm has features among the 11 by ignoring one features.

Conclusion

In this study, for developing the prototype knowledge is acquired using both manual and automatic knowledge acquisition method. The acquired knowledge is preprocessed using the Orange tool and through WEKA tool classification algorithms were comparatively analyzed using 10 folded cross-validation options and percentage splitting methods with six experiments and from six classification algorithms, JRIP algorithm with 10 folded cross-validation registers the highest results 94.16% accuracy. 39 rules were generated by JRIP and those JRIP rule were used to develop models that represents concepts and procedures involved in diagnosis and treatment of the stroke.

REFERENCES

Almadani, O. 2018. "Prediction of Stroke using Data Mining Classification," International Journal of Advanced Computer Science and Applications, vol. 9, no. 1.

Anusuya, R. 2013. “Fuzzy Membership Function Generation using DMS-PSO for the Diagnosis of Heart Disease”, Journal of Engineering Sciences Vol 4 Issue 2.

Anusuya, R. 2018. “An Investigation of Blood sugar level and predicting scrupulous insistence for diabetic recovery using data mining Techniques” CiiT International Journal of Artificial Intelligent Systems and Machine Learning, Vol 9, No 5, May - June.

Anusuya, R. 2013. “ A Novel Method to Diagnose Heart Disease Based on Genetic Algorithm and Fuzzy Decision Support System” Journal of Engineering Sciences Vol 4 Issue 2.

Anusuya, R. 2016. “A Novel k-Singular Value Decomposition Clustering Approach for Cancer Diagnosis”, Data Mining and Knowledge Engineering, Vol 8, No 10.

Bontia, R. A. B. 1985. "The enigma decline in stroke death in united state," vol. 27.

Cheritian, R. 2013. "Expert systen Based medical stroke prevention," *Journal of computer science*, no. 1549.

Engelmore, E. F. A. R. 1993. "Expert System and Artificial Intelligence”, Japanese," International journal of computer science and artificial intellgent.

Fayyad, 1996. "Data Mining and Knowledge discovery," oct.

Jain, D. R. 2014. Introduction to data mining techniques, India: IEEE.

Rajalakshmi, 2003. "Data mining techniques and Application.

Rajalakshmi, D. S. "Overview of Datamining techniques and Applicatiopn," Internashinal Journal of Science and Research.

Stroke Disease," (Online). Available: http://www. wekipdia.com. (Accessed 12 Dec 2018).

****