A Clinical Decision Support System for Identification of Factors Causing Stroke in Adults

Nonita Sharma, Ravi Sharma, Govind Singhal, Punit Sharma, Shouvik Banik

Abstract*: Among neurological patients, stroke is a significant concern that can lead to permanent disability or may cause death. Determining the contributing factors to stroke would better enable proactive forms of healthcare measures to be taken for reducing mortality and other effects of this disease. Numerous works have been carried out for determining the various factors which contribute to stroke. The goal of this work is to identify the factors which may cause stroke by proposing a Decision Support System. In this research work, various medical and psychological parameters are analyzed with the aim of determining association between these parameters and stroke. The methodology includes appropriate feature selection, attribute type conversion, univariate and multivariate analysis. Data mining techniques such as Chi square test and Frequency Distribution analysis are used for hypothesis testing and drawing inferences. The analysis shows that psychological factors such as Marital Status, Employment Status, and Age Group are the major contributors, which may lead to stroke in comparison to Medical factors. The findings can be used as a proactive measure for individuals who have high possibility of developing stroke. This work can also be used as a foundation to build a recommendation system for prevention of stroke.

Keywords: Correlation, Data mining, Decision Support Systems, Frequency Distribution, Healthcare, Stroke

I. INTRODUCTION

Stroke is caused by an unexpected change of blood flow in the brain. Many circumstances can cause a stroke, including abrupt blockage of blood flow, excessive bleeding of tissues in the brain when some blood vessel suddenly gets burst. These are referred to as Ischemic Stroke and Hemorrhagic Stroke. Transient Ischemic Attack or mini-stroke is when the blockage of blood flow is temporary and lasts for a very short time. A patient facing stroke needs immediate attention and treatment. American Stroke Association has given FAST (Face, Arms, Speech, Time) method to identify such situations [1]. Global Health Estimates 2016, a report of the World Health Organization states that Stroke and Ischemic heart disease are the top cause of deaths globally [2]. These two contributed 15.2 million deaths out of 56.9 million deaths globally. Moreover, these diseases were in the top 10 causes of death in the previous 15 years.

Revised Manuscript Received on January 15, 2020

Nonita Sharma, Department of Computer Science and Engineering, Dr. B. R. Ambedkar National Institute of Technology Jalandhar, Jalandhar, India.

Ravi Sharma, Department of Computer Science and Engineering, Dr. B. R. Ambedkar National Institute of Technology Jalandhar, Jalandhar, India.

Govind Singhal, Department of Computer Science and Engineering, Dr. B. R. Ambedkar National Institute of Technology Jalandhar, Jalandhar, India.

Punit Sharma, Department of Computer Science and Engineering, Dr. B. R. Ambedkar National Institute of Technology Jalandhar, Jalandhar, India.

Shouvik Banik, Department of Computer Science and Engineering, Dr. B. R. Ambedkar National Institute of Technology Jalandhar, Jalandhar, India.

A large number of studies in the domain of stroke data analysis have been carried out over a while. A wide variety of research work proposes machine learning algorithms such as Decision Tree, Bayesian Learning, and Artificial Neural Network to learn the symptoms and predict the presence of stroke disease [3]-[5]. The models like Neural Network-Based Expert Decision Support System for Stroke diagnosis are used to set the preliminary and final diagnoses to get the examination and treatment plans to analyze statistical data [6]. The literature includes predicting the probability of developing stroke based on various medical factors of an individual, such as heart disease, hypertension, and immunity disease [7]. Performance comparison analysis is also carried out to compare various approaches for prediction of different diseases [8][9]. Stroke is a severe disease that has caused millions of premature deaths and disability in the long term. The after-effect of the stroke may cause paralysis. Some studies propose the design of automatic devices to help the patient to recover faster [10].

The objective of this paper is to propose a decision support system based on data-mining techniques to identify factors that may lead to stroke [11]. In this work, a dataset of 43401 subjects is analyzed, their medical and psychological factors are reviewed, univariate and multivariate analysis is explored to get insights. Based on the analysis done in this work, it is found that subjects that faced stroke share similar psychological factors up to some extent. These psychological factors are their employment status, marital status, and Age group. One of the medical factors for which subjects share the same relation is Body Mass Index (BMI), and it is the only contributing unit out of all medical factors. Features such as heart disease, average glucose level, and hypertension have the least impact on the subjects who have suffered from the stroke.

II. PROPOSED METHODOLOGY

Unstructured and semi-structured problems are those in which human intuition is needed. A decision support system (DSS) helps decision makers to solve such problems by utilizing data [11]. The proposed methodology for construction of DSS for identification of factors causing stroke in adults is shown in Fig.1.

A. Data Selection

Selecting a subject-oriented data is always a crucial step in data analysis. Factors such as the number of subjects and dimensions directly affect the analysis.

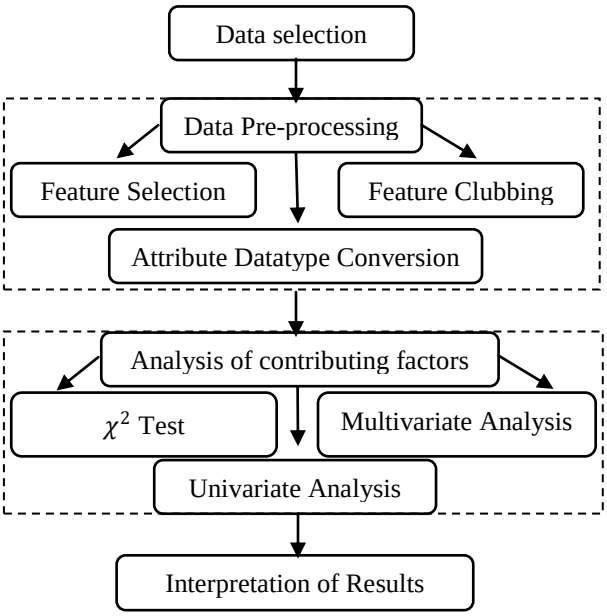

Fig. 1 Knowledge Discovery Steps in DSS Construction

B. Data Pre-processing

Data pre-processing is the next key step performed after data selection to convert the raw data into an intelligible format by cleaning, selecting the features of interest, and normalizing it. This also includes data attributes conversion when analysis requires converting one attribute type to another one [12]. Clubbing multiple attributes when logical interpretation is needed can also be done in this step; however, the clubbing does not affect the analysis but gives a better visualization.

C. Analysis of contributing factors

Co-relation analysis of the variables is evaluated to develop a better understanding of how strongly two attributes are related to each other. Chi-Square test of independence is used to find out the correlation between two parameters of categorical type. The null hypothesis is framed prior to performing this test, which attempts to state that there is no correlation between two variables, and an alternate hypothesis justifies the co-relation. The null hypothesis will be accepted only when chi-square test value is less than the chi-square table value else alternate hypothesis will be accepted.

$$\chi^2 = \sum_{i=1}^{m}\sum_{j=1}^{n}\frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}} \quad (1)$$

Chi-Square test of independence uses the formula mentioned in (1) where O_{ij} is observed frequency and E_{ij} is the expected frequency of the event. The value of E_{ij} is calculated as formula mentioned in (2) where n is the total number of data tuples, $count(A = a_i)$ is the frequency of tuple having a_i for attribute A and $count(B = b_j)$ is the number of tuples having value b_j for B [12].

$$E_{ij} = \frac{count(A = a_i) \times count(B = b_j)}{n} \quad (2)$$

The univariate analysis of the parameters portrays the distribution of the dataset. Statistical analysis such as mean, median, standard deviation, central tendency, and average obtained from frequency tabulated data makes data management and operations more effective and insightful [13].

Multi-variate analysis of the parameters in the selected dataset is evaluated to figure out the strength of association between the considered parameters and the target attribute considering more than one parameter at a time. This helps in getting a proper idea about the multiple parameters affecting the target attribute collectively [13].

D. Interpretation of Results

Interpretation of the result is done by testing the assumed hypothesis and interpreting the frequency distribution. Assumptions are re-evaluated in this step to generalize the results.

III. RESULT AND DISCUSSION

Dataset under consideration has 43401 records, out of those 783 individuals have suffered from stroke [14]. Each individual has 11 features describing their demographic characteristics and one target attribute which states whether an individual suffered from stroke or not.

Table- I: Data Dictionary

Feature	Description
id	Patient ID
gender	Gender of the patient
age	Age of the Patient
hypertension	0 – no hypertension, 1 – Suffering from hypertension
heart_disease	0 – no heart disease, 1 – suffering from heart disease
ever_married	Marital status
work_type	Type of Occupation
Residence_type	Type of residence (Urban/Rural)
avg_glucose_level	Glucose level measured after meal
bmi	Body mass index
smoking_status	Patient's smoking status

Table-I shows the demographic characteristics of an individual where *id* is unique identifier of the patient, *age* and *gender* are in their implicit meaning. *hypertension* indicates whether an individual is suffering from high blood pressure or not and *heart_disease* indicates any prior history related to heart disease. *ever_married* states marital status of an individual who may be divorced. In *work_type* Private, Self Employed, Government Job, and Children are possible classes which show their profession. Urban and rural are two possible residence type which is represented by *Residence_type*. Average glucose level of an individual measured after having meal is recorded in *avg_glucose_level* , *smoking_status* classifying the individuals based on formerly smoked, never smoked, and smokes class, and *bmi* is the body mass index of an individual which measures ratio of weight and height. *age*, *avg_glucose_level*and*bmi* are quantitative features rest all are categorical [14].

After selecting the appropriate dataset, the next step is to pre-process the data. In data pre-processing relevant features are selected. Patient id is not significant because it acts as an identifier of an individual. Except patient id, all other features are selected for further analysis. The selected features are clubbed together into medical factors and psychological factors based on their nature as shown below.

1. Medical factors

$$hypertension, avg\_glucose\_level, bmi, heart\_disease, smoking\ status$$

2. Psychological factors

$$age, gender, ever\_married, work\_type\ , Residence\_type$$

All the quantitative features have been converted into categorical features for co-relation, univariate, and multivariate analysis. As per the World Health Organization adults having BMI greater than 24.9 fall under the category of overweight and underweight when BMI is less than 18.5, therefore BMI is classified into 3 categories underweight, normal, and overweight [15]. Furthermore age is classified into four categories teenager, adulthood, late adulthood, and old age having age less than equal to 18, between 19 to 40 (inclusive), between 41 to 60 (inclusive), and greater than 60 respectively. Glucose level is classified into three categories low, Normal, and high having glucose level less than 70, between 70 to 130 (inclusive), and greater than 130 respectively [16].

Table-II: Chi Square test

Feature	Test Value	Table Value (0.05)	Test Output
gender	5.564	3.8415	Correlation
hypertension	246.293	3.8415	Correlation
heart_disease	561.683	3.8415	Correlation
ever_married	224.483	3.8415	Correlation
work_type	357.87	9.488	Correlation
residence_type	0.219	3.8415	No Correlation
smoking_status	34.655	5.991	Correlation
age	1188.592	7.815	Correlation
avg_glucose_level	251.355	5.991	Correlation
bmi	66.682	5.991	Correlation

To find the correlation between two features we have performed Chi square test of different attributes against target attribute *stroke* as shown in Table-II. In the above table, Feature represents the attribute which is tested for correlation against *stroke*. Second and third column represent the value of chi square test and the value from chi square table with significance value 0.05 respectively. The last column in the table shows the output of the test. Based on the test it is found that there is no correlation between *residence_type* and *stroke* . However, there is a possibility of correlation between rest of the attributes and *stroke* . Hence *residence_type* will not be considered for the further analysis.

The selected dataset is analyzed using Datapine platform [17]. To get better insights from the data, univariate and multivariate analysis is performed by considering only those individuals who have suffered from the stroke.

Table-III: Univariate Analysis Of Medical Factors

Features	Category	Stroke Count
hypertension	Hypertension	200
	No Hypertension	583
avg_glucose _level	Low	98
	Normal	395
	High	290
bmi	Underweight	8
	Normal	128
	Overweight	507
heart_disease	Heart disease	177
	No Heart disease	606
smoking_status	Formerly smoked	221
	Never smoked	284
	Smokes	133

Table-III shows the univariate analysis of medical factors. All the categories of the features are listed in the table with their corresponding stroke count. Some inferences which can be drawn from the table are that majority of individuals who suffered from stroke, never suffered from hypertension and heart disease, and they have normal glucose level. On the other hand they are overweight.

Table-IV: Univariate Analysis Of Psychological Factors

Features	Category	Stroke count
age	Teenager	2
	Adulthood	22
	Late Adulthood	193
	Old age	566
gender	Male	352
	Female	431
ever_married	Married	703
	Unmarried	80
work_type	Children	2
	Government Job	89
	Private	441
	Self Employed	251

It is evident from the Table-IV that a high frequency of individuals who suffered from stroke are old age people, married and either works in private sector or self-employed. On the other hand gender does not give a clear distinction on stroke count.

Table-V: Multivariate Analysis of Medical Factors

bmi	*hyper_ tension*	*heart_ disease*	*smoking_ status*	*avg_ glucose level*	*Stroke count*
Over weight	No	No	never smoked	Normal	58
Over weight	No	No	formerly smoked	Normal	44
Over weight	No	No	never smoked	High	37
Over weight	No	No	formerly smoked	High	28
Over weight	Yes	No	never smoked	Normal	25
Over weight	No	No	smokes	Normal	24

Normal	No	No	never smoked	Normal	21
Over weight	Yes	No	formerly smoked	Normal	21
Over weight	Yes	No	never smoked	High	21

Table-V shows all the medical factors with respect to the stroke count. Table is sorted in decreasing order of stroke count, only higher frequencies of stroke are considered and shown. All the medical factors are combined together for drawing insights in terms of decision rules as below.

$$if\ BMI = 'Overweight'\ AND\ heart\_disease = 'No'\ then\ Stroke\ Count > 250$$

Based on the table it can be concluded that majority of individuals who are overweight and doesn't suffer from hypertension and heart disease are in maximum stroke count.

Table-VI: Multivariate Analysis of Psychological Factors

age	*work_type*	*ever_ married*	*gender*	*Stroke count*
Old age	Private	Yes	Female	70
Old age	Private	Yes	Female	70
Old age	Private	Yes	Male	68
Old age	Private	Yes	Male	67
Old age	Self-employed	Yes	Female	61
Old age	Self-employed	Yes	Male	50
Old age	Self-employed	Yes	Female	46
Old age	Self-employed	Yes	Male	39
Late adulthood	Private	Yes	Female	32
Late adulthood	Private	Yes	Female	27
Late adulthood	Private	Yes	Male	27
Late adulthood	Private	Yes	Male	23
Old age	Govt_job	Yes	Female	20
Old age	Govt_job	Yes	Female	17
Old age	Private		Female	12
Late adulthood	Govt_job	Yes	Female	11

All the psychological factors are listed in Table-VI and multivariate analysis is performed. Table is sorted in descending order of stroke count. The table generalizes below decision rule.

$$If\ age = 'Old\ Age'\ and\ (work\_type = 'Private'\ Or\ work\_type = 'Self\_employed')\ and\ ever\_married = 'Yes'\ then\ Stroke\ count > 470$$

Based on the decision rule it can be inferred that individuals who are of old age, married and either working in private sector or self-employed are in highest stroke count.

IV. CONCLUSION AND FUTURE SCOPE

In this work a clinical decision support system is proposed for identification of factors which may cause stroke in adults. The manuscript consist of univariate and multivariate analysis which shows that majority of people who suffered from stroke have normal average glucose level, no hypertension, no heart disease, and they do not smoke but they are overweight. On the other hand analysis of psychological condition shows that people who are of old age, married, and either self-employed or working in private sector are more in numbers who have suffered from stroke. However gender is least contributing psychological factor. In conclusion, psychological factors are the major contributor to the individuals who have suffered from the stroke compared to given medical factors except BMI.

Above work can be further extended by proposing a predictive or probabilistic model which can predict the possibility of stroke for an individual based on their demographic characteristics.

REFERENCES

1. American Stroke Association. Available: https://www.stroke.org/
2. Global Health Estimates 2016. Available: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death
3. Sudha, A., Gayathri, P., & Jaisankar, N. (2012). Effective analysis and predictive model of stroke disease using classification methods. International Journal of Computer Applications, 43(14), 26-31.
4. Amini, L., et al. (2013). Prediction and control of stroke by data mining. International journal of preventive medicine, 4(Suppl 2), S245.
5. Almadani, O., and Alshammari, R. (2018). Prediction of Stroke using Data Mining Classification Techniques. International Journal of Advanced Computer Science and Applications, 9(1), 457-460.
6. Kupershtein, L. M., et al. (2017). Neural expert decision support system for stroke diagnosis. In Photonics Applications in Astronomy, Communications, Industry, and High Energy Physics Experiments 2017 (Vol. 10445, p. 104453I). International Society for Optics and Photonics.
7. Colak, C., Karaman, E., & Turtay, M. G. (2015). Application of knowledge discovery process on the prediction of stroke. Computer methods and programs in biomedicine, 119(3), 181-185.
8. Sharma, N., & Juneja, A. (2019). Extreme Gradient Boosting with Squared Logistic Loss Function. In Machine Intelligence and Signal Analysis (pp. 313-322). Springer, Singapore.
9. Sultana, N., & Sharma, N. (2018, December). Statistical Models for Predicting Swine F1u Incidences in India. In 2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC) (pp. 134-138). IEEE.
10. Aminuddin M., Tarmizi A. (2019), Automatic Vibration Device for Stroke Patients International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, 8(2), 2329-2332.
11. Efraim Turban, Jay E. Aronson, Ting-Peng Liang. Decision Support Systems and Intelligent Systems. 7th ed. Prentice-Hall, Inc. Upper Saddle River, NJ, USA; 2004.
12. Han J, Kamber M. Data Mining: Concept and Techniques. 2nd ed. California: Morgan Kaufmann Publishers; 2006.

13. Barbara G. Tabachnick, Linda S. Fidell. Using Multivariate Statistics. 7th ed.
Pearson, 330 Hudson Street,NY NY 10013
14. Kaggle (Data Source). Available: https://www.kaggle.com/asaumya/healthcare-dataset-stroke-data#test_2v.csv
15. World Health Organization Chart of BMI categorization. Available: http://www.euro.who.int/en/health-topics/disease-prevention/nutrition/a-healthy-lifestyle/body-mass-index-bmi
16. American Diabetes Association. Available: https://www.diabetes.org
17. Datapine Data visualization and Business Intelligence Tool. Available: https://www.datapine.com/

AUTHORS PROFILE

Dr. Nonita Sharma is currently working as Assistant Professor in the Department of Computer Science & Engineering at Dr. B.R. Ambedkar National Institute of Technology Jalandhar. Her research interests are Data Mining, Wireless Sensor Network, Blockchain Technology, Internet of Things (IoT). She has published more than 20 papers in peer reviewed journal and international conferences. She has published a book entitled Extreme Gradient Forecasting for Data Mining Applications. She has received Best Paper Award in the International Conference held at CDAC Mohali, India.

Ravi Sharma completed his post-graduation in Computer Science & Engineering from Himachal Pradesh University, Shimla. He is a research scholar in Dr. B.R. Ambedkar National Institute of Technology Jalandhar, Punjab in Department of Computer Science & Engineering.

Govind Singhal completed his graduation in Computer Science & Engineering from Vellore Institute of Technology Vellore, Tamil Nadu. He is pursuing masters in Computer Science & Engineering from Dr. B.R. Ambedkar National Institute of Technology Jalandhar, Punjab

Punit Sharma graduated from Dr. A.P.J. Abdul Kalam Technical University, Uttar Pradesh in Computer Science & Engineering. He is pursuing masters in Computer Science & Engineering from Dr. B.R. Ambedkar National Institute of Technology Jalandhar, Punjab.

Shouvik Banik graduated from Jalpaiguri Government Engineering College, West Bengal in Information Technology. He is pursuing masters in Computer Science & Engineering from Dr. B.R. Ambedkar National Institute of Technology Jalandhar, Punjab.