**1.** **Abstract:**

At present, stroke has become a common disease. One in three people die within a year of having a stroke. Almost one in five people who experience a stroke are under the age of 55. Men are more likely to suffer a stroke and at a younger age. Stroke kills more women than breast cancer. A stroke is a serious medical condition that requires emergence care. A stroke can cause lasting brain damage, long-term disability, or even death.

A stroke may affect only one side of the body or part of one side. It can cause paralysis (an inability to move) or muscle weakness, which can put you risk for falling.

In health care, the analytical data mining algorithms have great impact on building the machine learning models to carry out disease prediction. As a stroke can cause lasting brain damage, it is a very serious issue, so an earlier prediction of the risk of stroke may help the treatment as a strong guidance. Considering this issue using the data mining techniques such as classification, clustering, association rules and using various types of software tools like WEKA software tool, a proper dataset should be trained and analyzed. In this paper our main objective is to construct an appropriate dataset for this thesis and provide an improved datamining classification based architecture to predict the risk of this diseases by analyzing the clinical data of problems.

**2. General Term:**

Data Mining and techniques.

**3.Keywords:**

Ischemic Stroke, Machine Learning Algorithm, Support Vector Machin, Naïve Bayes, Artificial Neural Network, analytical medical data.

**4.Introduction:**

A stroke happens when blood flow is occluded in a part of the brain. The lack of oxygen damages the brain cells that can have potentially disabling effects on the patient. When it happens the brain cell divested of oxygen and brain cells begin to die. Then the abilities controlled by that area of the brain such as memory and muscle control are lost. Some strokes affects the muscles used to urinate. There are two types of stroke. They are:

1. Ischemic Stroke
2. Hemorrhagic Stroke

**1. Ischemic Stroke**

An Ischemic stroke is the most common type. It occurs when an blood supply is cut off to part of the brain. It accounts for the majority of all strokes. Approximately 85% of strokes are ischemic caused by vascular occlusion.

An ischemic can occur because of lesions caused by atherosclerosis. These lesions may form in the small arteries of the brain and they can block blood flow to the brain.

**2. Hemorrhagic Stroke**

A hemorrhagic stroke is a different kind of stroke caused by bleeding in the brain. It happens when a blood vessel breaks and bleeds into the brain. Blood spills into or around the brain and creates swelling and pressure, damages cells and tissue in the brain.

Hemorrhagic stroke is caused by a rupture in a weakened blood vessel in the brain. Hemorrhagic stroke account for about 20% of all strokes.

In this thesis paper we will predict the risk of stroke. A man who has not yet stroked, will talk about the possibility of having stroke in the future. It is a more challenging task in healthcare sectors to predict the diseases from the voluminous medical databases. At present, data mining techniques will help us a lot to predict risk. Data mining techniques which includes classifications, clustering, association rule mining for finding risk prediction. In this research work Naïve Bayes Support Vector Machine (SVM) classifier algorithm are used for stroke risk prediction.

The brain is one of the largest and most complex organs in the human body. It is made up of more than 100 billion nerves that communicate in trillions of connections called synapses. [1] It is the central organ of the human nervous system. It controls most of the activities of the body, processing, integrating and coordinating the information it receives from the sense organs and making decisions as to the instructions sent to the rest of the body. Because of a ischemic stroke, brain cells start to die or damage. So symptoms occur in the body parts (face, eyes, arms, legs etc.) that these brain cells control. There are number of factors which increase the risk of having a ischemic stroke. Some of them are listed below:

* Age and gender
* High blood pressure
* Diabetes
* Ischemic heart diseases
* Smoking
* Family history of stroke
* Stress and depression
* Overweight and obesity
* Abnormal cholesterol levels

The remaining portion of the paper is organized as follows,

**5. Literature review:**

In this section reviews various research works that are related to the proposed work.

**Dr.S.Vijayarani**, **Mr.S.Dhayanand** have predicted about liver disease prediction using SVM and Naïve Bayes algorithms. They have made their data set within five hundred and seventy six instance and ten attributes. They implemented their works in Matlab 2013 tool. After doing their work they showed a result of an experiment that the accuracy of the performance of SVM algorithm is better than Naive Bayes algorithm.

**Pragati Agrawal and Amit Kumar Dewangan** have gave a brief survey on the techniques used for the diagnosis of diabetes –mellitus. They described in their paper about diabetes. They described in their paper about diabetes. They designed their data set on various classification algorithms like SVM, KNN, Naïve Bayes, ID3, C4.5, C5.0 and CART to classify the diabetes data. They have made a comparison of accuracy of these models and they showed that SVM gives best accuracy of classification as 81.77% compare to others.

**Ebrahim Edriuss et.al** clarify the modeling of breast cancer as classification task and describes the implementation of Neural network(NN) and Support Vector Machine approach for classifying breast cancer is either petit or malignant. Dr William H.Wolberg created a dataset to diagnosis breast cancer consist of 400 observations of patients with breast cancer among which 300 are petit band 100 are malignant status. They has 20 features. The work has done in two experiment in which first one is done by using SVM open source tool for multi class SVM, which uses Crammer and singer method. The 2nd one is done by busing Neural Network. The result of the performance and accuracy of both NN and SVM algorithm were compared. In this paper there are shown that the NN technique is more efficient compared to SVM technique in breast cancer detection.

**Dr.D.Asir Antony et.al** conducted a prediction of diabetes using medical data. In his research paper they showed a prediction of diabetes by using WEKA software with the configuration of computer system 4GB RAM, Intel(R), Core(TM)2,CPU 1.73GHz Processor, Windows 7,64-bit operating system. For preparing this research paper ,the data have been collected from University of California, Irvine and in this paper different machine learning algorithms( Naïve Bayes(NB),Multilayer Perceptron(MLP),decision tree-based random forests(RF) )are used to build the model and those model is tested with different testing methods such as FCV(10-fold cross Validation), PS(Percentage Split) and UTD(Use Training Dataset) to evaluate the accuracy of the model. It is observed that the technique with pre-processing(WPP) increase the accuracy of the machine learning algorithm more than the technique without pre-processing (WOPP)

**Joseph A. Cruz et.al** presented a paper about the applications of Machine learning in cancer prediction and prognosis. In their paper they showed that different types of machine learning algorithm can be used to predict different types of cancer and they showed a histogram to show how many papers are published during years nearly 1994-2005 to predict cancer risk, recurrence and outcome. They have attempted to explain, compare and assess the performance of machine learning algorithms in this field and they also clarify that machine learning methods generally improve the performance or predictive accuracy of most prognosis, especially when compared to conventional statistical or expert-based systems.

**Mike E Janicek, MD et.al** discussed about Cervical Cancer to prevent, diagnosis and therapeutics. They discussed what the Cervical cancer is, what the risk factors will be occur because of Cervical cancer. They described about HPV, types of HPV, role of HPV and its types, Us Epidemiology, Liquid-Based Sampling Techniques and much more. They told in their paper that surgical, radio therapeutic, and chemo radio therapy approaches comprise the successful treatment modalities for invasive cervical carcinoma.

**6. Methodology**

**7. Architecture**

**8. Necessary tools**

**9. Conclusion**

**10. References**