Literature Review #3: “Identifying Stroke Indicators Using Rough Sets”

The paper we will discuss is “Identifying Stroke Indicators using Rough Sets,” which was published on November 19th, 2020, and written by Soumyabrata Dev. The number two leading cause of death is dying from a stroke. In this article, they talked about finding different indicators of a stroke. Many different tactics have been used for data mining to predict the leading stroke indicators. The data was taken from electronic health records, which contain thousands of different features. Allowing feature selection can help make a more accurate prediction. After using the proposed technique on public electronic health records, it was found that age, average glucose level, heart disease, and hypertension were the essential way attributes in detecting strokes in patients.

According to the World Health Organization, the estimated amount of deaths from cardiovascular diseases was 17.7 million in 2017, and 6.7 million of them were from having a stroke. There is a world stroke day that is observed every year where people are being educated on how strokes can be prevented on October 29 every year.

Machine Learning (ML) has been created to be applied to electronic health records to help predict the risk of having a stroke in each patient it is used on. The way that the ML predicts is by using feature selection of some features in the electronic health records. The most important question that is asked when researching this is how we can select the most discriminatory factors from a large electronic health record that has thousands of different features. The only way to predict the chance of having a stroke efficiently is by using the core risk factors that are very well associated with a stroke. The way that the article summarized the process of their work was by first identifying the critical features in electronic health records to detect a stroke. Second proposing a novel rough set that can be used in another set of feature data in addition to the original. Third, release the open-source code to the research community.

The most important thing for detecting a stroke is to make sure they select the features that will have the most influence when it comes to the stroke's outcome. There is a method in the article that is discussed that used 9 critical features out of 28 features. Using models like support vector machines, decision trees, ensemble methods & deep neural networks. Another method created that was used was the Support Vector Machine (SVM) which was used to find the most critical risk factors in the dataset. Using this technique, it was found that the 6 most effective features should be used as stroke indicators, which were age, high blood pressure, serum creatinine, lactate dehydrogenase, and alkaline phosphate.

Moving forward, the paper begins to describe how they used the rough set theory created by Pawlak. It’s a statistical approach to analysis and data mining applied in healthcare, finance, manufacturing, finance, engineering, and other fields. Rough sets pull a subset of attributes that pertain to the original attributes. Making it easier to use feature selection or attribute selection on the set instead of directly on the data itself.

There have been many techniques created thru the use of rough sets like genetic algorithms and multi-granulation methods. Roughsets are going to be more suitable for datasets that have numerical features & it won't work on datasets that have binary features. There is a way when modifying the rough set technique for it to work on a binary versions of datasets.

The article proposes a modified rough set model to deal with binary attributes in datasets. Traditional set-theory techniques do not consider binary features, but this proposed approach uses the relevance factor to identify the most discriminatory attributes. The retention index is defined to determine the ratio of elements in the conditional attribute set that are definite members of the decision attribute.

In this study, they evaluated the efficiency of 10 patient attributes in detecting stroke and electronic health records. The attributes include age, gender, marital status, work type, residence type, heart disease condition, body mass index, smoking status, glucose level, and hypertension status. We conducted experiments using the Rough Set Theory approach on the discretized dataset. The results showed that age, hypertension status, and glucose level were the three most significant attributes in stroke detection.The results indicate that age is a significant factor in stroke detection. Older patients have a higher risk of stroke. The hypertension status of the patient is a significant factor in stroke detection. Patients who have hypertension are more likely to have a stroke. The average glucose level is also an essential attribute in indicating a stroke. Patients who have a high glucose level are more likely to have a stroke. The results from the study demonstrate the potential use of electronic health records to detect strokes. The study shows that age, hypertension, status, and glucose level are the three most significant factors in detecting a stroke. These demonstrations are consistent with previous studies that have identified these factors as necessary for detecting a stroke. The performance of a prediction model for stroke detection can be affected by the amount of data in a database. The proposed method was tested by dividing the dataset into 10 different fractions, and the results showed that there is not a significant variance in the correlation values. The relevance value is high when the database fraction is at 70% and 100%, which indicates that the amount of data and knowledge in the databases can heavily affect the results.

In summary, the paper's authors proposed an efficient feature selection technique based on rough set theory to identify critical stroke indicators from a large medical dataset containing 29,072 patient records with 10 different attributes. The proposed technique was modified to work with binary feature sets and was compared with other popular feature selection techniques. The results showed that age, heart disease, hypertension, and average glucose level were the most critical attributes for stroke detection. The authors plan to use the identified features to develop a machine-learning framework for improved stroke detection and investigate the use of their proposed rough-set technique in other applications. The study highlights the

importance of feature selection in stroke detection and the effectiveness of rough set theory for this purpose.