AIRO JOURNALS

Free / Unpaid
Peer Reviewed
Multidisciplinary
International

ISSN:2320-3714
Volume:3 Issue:3
July 2021
Impact Factor:5.7
Subject:Computer Science

A THEORETICAL STUDY ON DATA MINING IN HEALTHCARE SECTOR

Swapna Bhavsar
Research Scholar
University of technology, Jaipur
Dr. Ashish Chaurasia,
Professor
University of Technology, Jaipur

DECLARATION: I AS AN AUTHOR OF THIS PAPER / ARTICLE, HEREBY DECLARE THAT THE PAPER SUBMITTEDBY ME FOR PUBLICATION IN THE JOURNAL IS COMPLETELY MY OWN GENUINE PAPER. IF ANY ISSUE REGARDING COPYRIGHT /PATENT/ OTHER REAL AUTHOR ARISES, THE PUBLISHERWILLNOT BE LEGALLYRESPONSIBLE. IF ANY OF SUCH MATTERS OCCUR PUBLISHER MAY REMOVE MY CONTENT FROM THEJOURNALWEBSITE. FOR THE REASON OF CONTENT AMENDMENT/OR ANY TECHNICAL ISSUE WITHNO VISIBILITY ON WEBSITE/UPDATES, I HAVE RESUBMITTED THIS PAPER FOR THE PUBLICATION. FOR ANYPUBLICATION MATTERS OR ANY INFORMATIONINTENTIONALLY HIDDEN BY ME OR OTHER WISE, I SHALL BE LEGALLYRESPONSIBLE. (COMPLETEDECLARATION OF THE AUTHOR AT THE LAST PAGE OF THIS PAPER/ARTICLE

ABSTRACT

Data mining is a nontrivial process that involves extracting hidden patterns from massive amounts of data. It is a rapidly growing discipline that makes use of statistics, visualisation, machine learning, and other data manipulation and knowledge extraction techniques in order to get an understanding of the relationships and patterns that are concealed within the data. In this era of the internet and mobile devices, the volume of data is expanding at an enormous rate, and as a result, analytics has become the buzz word of today's market in both the IT and non-IT businesses. The majority of potential applications for data mining and machine learning techniques are being explored in the medical field. This is because large data presents a number of issues that need to be addressed. Due to the fact that the pace of rise in patients is linked to the rate of population growth and the changes in lifestyle, there is a high need for data processing in the healthcare industry.

Keywords: Data Mining, Health, Machine Learning.

1. INTRODUCTION

The advent of the internet and other technological advances has ushered in a new era of digital connectivity, which links individuals and their many electronic tools. Real-time data capture is made possible by recent technological advancements in conjunction with distributed computing. This encompasses not only digital material and transactions, but also the real-time networking of devices and appliances (also known as internet of things). The term "internet of things" refers to the expansion of the internet that will occur as more and more tangible items (such as consumer electronics and physical assets) are connected to the network. These include things like payments (through mobile and Near Field Communication), recognition (picture, voice, etc.), and sensors, among other things (embedded). Because of this connectedness, a concept known as context-aware computing (CAC) has emerged. CAC makes use of information about a user that is found in the environment in order to enhance the quality of the interaction and provide context aware services (has applications across domains). The expansion and further development of the internet, in conjunction with the widespread availability of various communication technologies, is fast reshaping the foundation upon which our lives are built.

2. APPLICATIONS OF DATA MINING IN HEALTHCARE SECTOR

In this day and age, medical industries have a greater need for data mining than ever before because these industries are producing an increasing amount of data. It has been suggested that data mining is moving from being a desirable to a necessary part of the healthcare industry rather than remaining a desirable part. The mining of data is helping to cut down on fraud in medical insurance. Applications for data mining can be utilised in a variety of healthcare settings, including hospitals and medical research facilities. One of the most important considerations in selecting a healthcare data mining technique is the kind of information that will be mined. The following are some of the ways that data mining programmes tackled the medical challenges.

- Classifying patients into categories
- determining the association between diseases and symptoms
- locating patients who have recurrent health issues
- identifying frequent patients and the nature of their condition
- Providing an estimate of the clinical diagnosis
- The study of medicine

In the field of healthcare, the massive amounts of data that are generated by the exchanges of clinical data are extremely challenging to process. The traditional method of analysis will not do when faced with these kinds of facts. Data mining is the solution to any and all concerns concerning medicine that have been posed by both professionals in the field and patients. Also, data mining analyses these massive amounts of data in such a way that the financial load has raised the necessity for decision-making by medical organisations based on the analysis of medical and economic data. Data mining is becoming increasingly popular in the field of healthcare for a number of reasons, one of which being the development of decision support systems and non-invasive diagnostic methods for diseases that have an alarming propensity to spread rapidly. Patients often find that some laboratory investigations, as well as the diagnostic treatments that follow them, are not only expensive but also intrusive and uncomfortable. It is an extremely uncomfortable process, for instance, when a biopsy is performed on a woman in order to diagnose cervical cancer. In order to accomplish this objective, an algorithm such as K-means clustering might be utilised to examine patients with cervical cancer. The clustering of data might produce a superior result on the basis of prediction as opposed to the opinion of a currently practising medical doctor.

There are a wide variety of different applications in the healthcare industry that could potentially benefit from data mining. It makes it possible to describe the activities of patients and view upcoming hospital appointments. It is helpful in identifying the patterns of effective medical treatments that have been given to patients for a variety of illnesses. There is a continuous improvement in applications for data mining in a variety of industries, which may supply more out of sight knowledge that boosts the effectiveness of the business and contributes to the expansion of the businesses.

3. ROLE OF DATA MINING IN HEALTH CARE

A cost-effective healthcare system that is designed to promote a community's or population's health requirements and is financed either directly by the government or by a committee that has been approved by the government is what is referred to as public healthcare. The public health system guarantees the financial stability of a healthy society. Instead of focusing on patients with character flaws, doctors in the public healthcare system deliver treatment to the entire community using equitable practises, which helps to ensure that society is a positive place to live. People are able to access more and better healthcare facilities for things like blocked sinuses, sore throats, or a broken limb at a significantly lower cost than other healthcare systems. Examples of these kind of conditions include: The hospitals and clinical facilities that are operated by the government-funded health care organisation are conveniently located for the populations. The health of the population as a whole, as opposed to the health of individual persons, is the focus of public health, which necessitates a collaborative effort and addresses disease prevention, treatment, and care from the point of view of the community's occupants. Whereas health care administration has focused internally on the organisation of health services, public health has looked outwards in the direction of the provision of quality care. The dual responsibilities of examining the health of the government and supporting counteractive activity each have a place for every member of society. Access to medical treatment is one of the most essential aspects of public health, and developing nations have a number of challenges in this area. A healthy population is largely dependent on having this access. Electronic health records are rapidly becoming an increasing number of organisations' standards in the health care industry. Because of the improved access they have to a substantial amount of patient information, healthcare firms are currently in a position where data mining can assist them in improving both the effectiveness and the calibre of their operations. Companies have also used information processing as fraud detection and a lot since the 1990s for operations such as loan marking. In today's day and age, in addition to prophetic analytics, many healthcare firms all across the world are also starting to understand the potential rewards of medical data processing. When information is processed, the goal is frequently to discover beneficial and understandable patterns by analysing extensive knowledge sets. This is true whether the information is being processed for business or for medical purposes. These information models help in predicting future market or information patterns, which in turn makes it easier to choose what actions to take in response to those trends.

In particular, information mining in the healthcare industry can significantly cut costs by increasing efficiencies, enhancing the quality of life of patients, and possibly even more substantially assisting in the process of saving the lives of many more patients. These benefits can be achieved through the company's investment in technology. The term "data processing" might mean many different things to very different people, such as "analytics" and "business intelligence." The most fundamental principle underlying data mining is the examination of large information sets in order to recognise patterns and then employing those patterns in order to forecast or estimate the likelihood of future events.

4. DATA MINING ALGORITHMS IN HEALTHCARE

The diagnosis, treatment, and prevention of disease, injury, and other physical and mental limitations in humans are all aspects of healthcare that are comprehensively covered by this field. In most nations, the sector of medicine and healthcare is through a period of fast change. A large amount of data, such as electronic medical records, administrative reports, and other benchmarking findings, is produced within the healthcare industry, which is one of the reasons why this sector is considered to be a rich source of data. However, these healthcare data are not being leveraged to their full potential. The process of data mining allows for the discovery of new and useful information hidden within these massive datasets. In the field of medicine, data mining is most commonly applied to the prediction of diseases as well as the provision of diagnostic assistance to physicians throughout the process of making therapeutic judgments. The following is a discussion of the many approaches that are utilised in the medical industry nowadays.

- **Anomaly Detection**

The most significant changes in an information set can typically be identified through the detection of anomalies in the data. In particular, three distinct anomaly detection methods, including normal support vector data description, density induced support vector data description, and Gaussian mixture, were utilised in order to evaluate the precision of anomaly detection on the unsure dataset of liver disorder acquired from UCI. The technique's accuracy is evaluated using the AUC precision. The uncertain dataset is quite likely to be present in all datasets; one solution to the problem would be to look for abnormalities in the data. It is very possible that the unreliable dataset may be accessed in all of the datasets; locating anomalies would be an excellent method for resolving the issue.

- **Clustering**

The task of identifying a limited number of categories or groups, known as clusters, as a means of describing the data is referred to as clustering. In the clustering approach to advanced medicine, the method of vector quantization was utilised to make predictions on patients' subsequent hospitalizations. The vector quantization device utilizes use of the algorithms X-means, K-means, and K-medoids. The clinical approach and the findings of the patient's laboratory tests were the sources for the data sets that were utilised in this investigation. The evaluation for each of the algorithms is carried out by utilising the Davies-Bouldin Index as the tests that Davies (1979) conducted. The K-means algorithm produced the best results, while the x-means algorithm produced fair results, and the K-medoids algorithm produced the worst results. The findings of these researchers provide a useful outcome that can assist in characterising the various types of patients who have a larger risk of being readmitted. Considering that this is the sole document that discusses the vector quantization, it is hard to make a contrast of the technique that is more significant.

- **Classification**

Classification is the process of discovering a predictive characteristic of learning that places a data item into one of several predetermined classes. This process takes place when a data item is classified.

Statistical

Both the Mahalanobis distance (MD) and the Mahalanobis room (MS) are utilised in order to generate statistical decisions that differentiate between one cluster and another. Additionally, the Mahalanobis room (MS) is utilised in order to represent the extent to which the recognised reference group observations are abnormal. The Mahalanobis Taguchi System, abbreviated as MTS, was employed by statistical classifiers in the process of developing a predictive model for pressure ulcers. Problems with class imbalance frequently occur in datasets used in healthcare. When employing information sets that are skewed or imbalanced, the use of data mining algorithms is frequently impacted by the skewness of the distribution. This problem frequently results in the tendency to produce extremely predictive classification precision over the majority class while producing poor precision over the minority class. [Cause and effect] Having such a nature to differentiate the degree of observed abnormality, this technique would be an effective technique for testing the data set by the information provided if it were applied. Additionally, this method is utilised while the MD is scaled in a suitable manner. The performance of the measurement scale with regard to this method is dependent on the vast gap that exists between typical and out of the ordinary occurrences. In order to be considered an algorithm that is appropriate for scaling, the MTS must display a better sensitivity and values than g-means does during the testing phase. The sensitivity and efficiency of the MTS have both been increased.

- **Discriminant Analysis**

In a discriminating assessment, the Linear Discriminant Analysis (LDA) technique is frequently utilised to estimate the class based on a set of measures that are derived from recent unlabeled observations. This technique is utilised by researchers all over the world. The LDA method was utilised by Duda R., P. and Hartand, D. Stork in order to predict the severity staging of Parkinson's disease patients via evaluations of nonmotor symptoms. The purpose of this study is to carry out a quantitative analysis of the internal relations that exist between the symptoms of the motor and those of the non-motor. The linear discriminant analysis is a conditional probability density function that is used by a predictor. This function adheres to a standard distribution and is determined by the class value that is given. The capability of the algorithm to incorporate statistical relationships among the predictor variables suggests that this algorithm would be suitable for exploring the linear constraint of this study in order to find the energy that is shared between the motor and the non-motor symptoms.

- **Decision Tree**

Many different study have looked into the effectiveness of using the decision tree technique for assessing clinical data. In its core, it consists of analysing the data and making use of the tree and the principles that govern it in order to make predictions about a dataset. Because it repeatedly separates observations into branches to form a tree, the decision tree cannot be used to provide predictive choices to correct imbalanced issues. This is because the decision tree is used to build trees. The choice tree is a structure that is similar to a tree and is composed of inner nodes, branches, and leaf nodes. In this structure, each branch represents a value associated with an attribute, each inner point contains a test on an attribute that was used for, and a leaf node represents the courses or class distributions that were predicted. The root node serves as the starting point for the classification process, which then moves its way along the tree based on the accuracy of each attribute's prediction. In order to build trimmed decision trees, the process includes information partitioning, the classification of information, the selection of decision tree categories, and the request to decrease fault trimming. Testing can be done with or without a vote, depending on how the information is partitioned. The Gini Index, Improvement of Information, and Gain Ratio are all examples of different types of Decision Trees. Last but not least, decreasing the amount of mistake trimming is useful in order to create more closed decision recommendations.

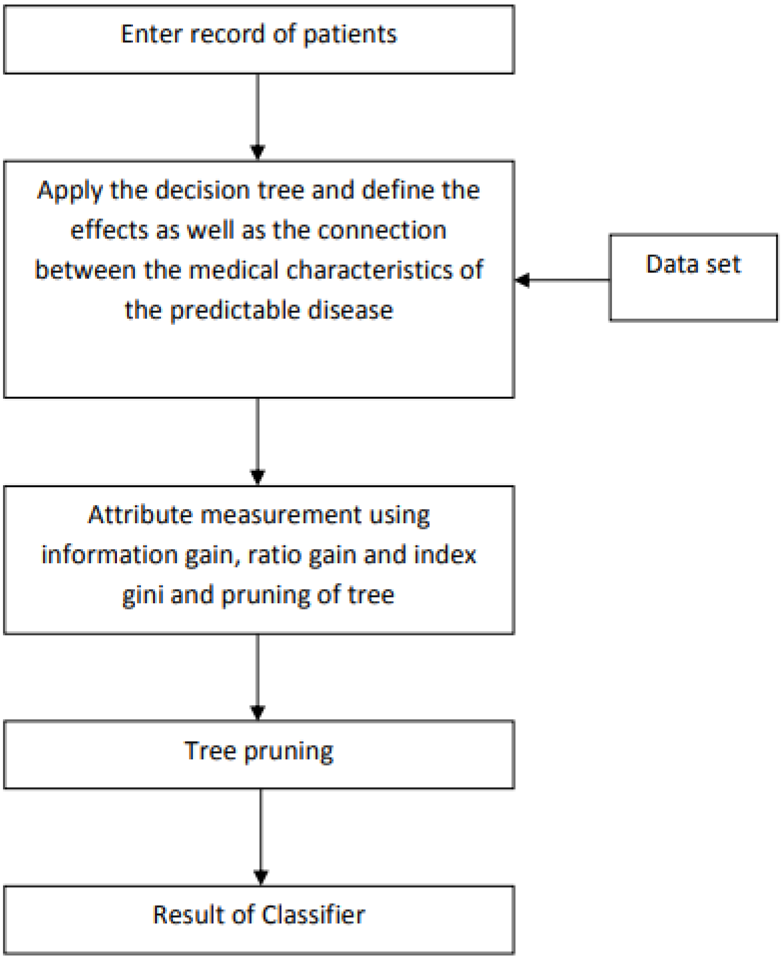
Enter record of patients
Apply the decision tree and define the effects as well as the connection between the medical characteristics of the predictable disease
Data set
Attribute measurement using information gain, ratio gain and index gini and pruning of tree
Tree pruning
Result of Classifier

Figure 1 Implementation of ID3 algorithm on patient data

- **Swarm Intelligence**

The Particle Swarm Optimization (PSO) technique is able to effectively identify the best or near-optimal solutions in large search spaces. The categorization approach will be both quicker and more accurate if it uses fewer characteristics. The PSO-based technique shows that it can improve general classification results since it uses PSO to choose optimal parameters in the classifiers that are being considered.

- **K-Nearest Neighbor**

An approach for classifying data that is based on an instance is called the k-nearest neighbour. In this approach, the parameter units are the samples that are used, and the underlying assumption of this algorithm is that the points in n-dimensional space are connected to all of

ISSN:2320-3714
Volume:3 Issue:3
July 2021
Impact Factor:5.7
Subject:Computer Science

the instances. Because the information is never lost in the training data, the technique has a lot of practical applications. This algorithm is very moment when processing each of the samples in the training set while classifying new data, and this procedure needs longer classification time. However, if the training data set is large, this algorithm would be suitable because it would be appropriate in this situation. When it comes to medical diagnostics, having precise classification systems is absolutely necessary.

- **Logistic Regression**

Logistic regression, often known as LR, is a method that can either be continuous, discrete, or a mixture of both types in addition to the binary goal by utilising the features that have been provided. After that, the LR performs certain calculations to come up with a linear input combination, which it then feeds into the logistic function. This strategy is common because it is easy to implement and it produces results that are comparable to those of other methods.

- **Bayesian Classifier**

The Bayesian classifier is capable of dealing with missing information in a natural and effective manner and has a high level of effectiveness in terms of computational efficiency. The Bayesian classifier also demonstrates by getting the models applied that the model is suitable as the average strategy has resulted in enhanced forecast precision and enables writers to extract more qualities from the information without being over-fitted. This is done by demonstrating that the model is suitable by getting the models applied. Bayesian classifiers provide predictions about the likelihood of class affiliation in such a way that they can determine the statistical likelihood that a given sample belongs to a particular class. In light of the observation, the Bayes theorem was applied to calculate the likelihood that a proposed diagnosis is accurate. The Naive Bayes model specifies the physical attributes and personality traits of a disease patient. It offers the possibility of defining a property for the anticipated state of each input. The Naive Bayes algorithm's implementation is depicted in Figure 2 for your reference.

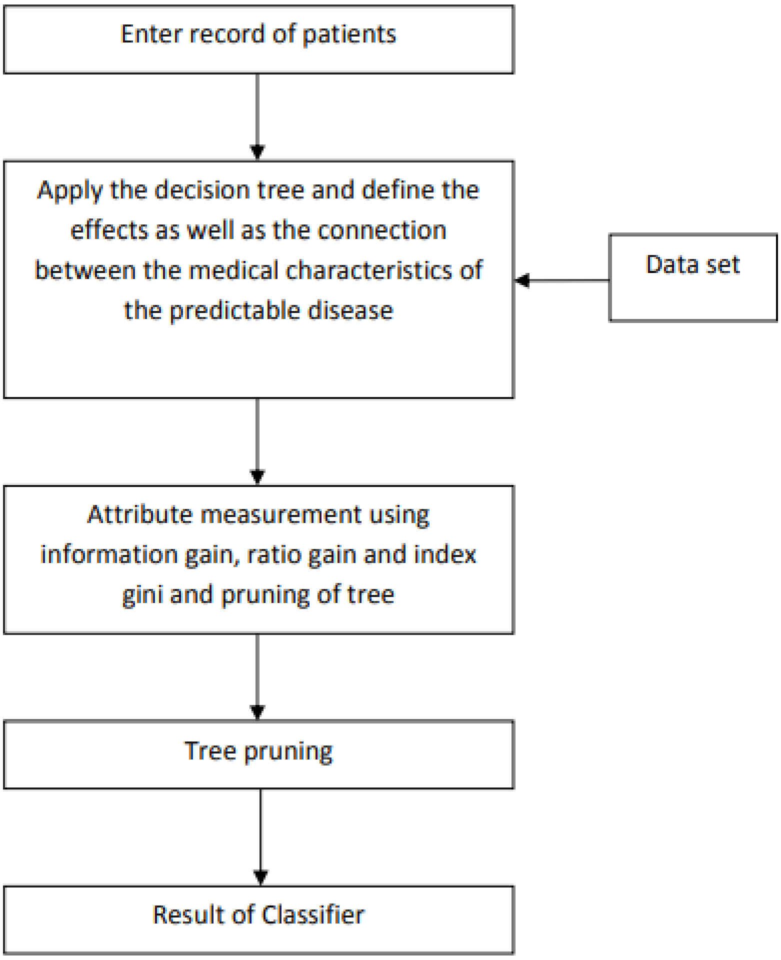
Enter record of patients
Apply the decision tree and define the effects as well as the connection between the medical characteristics of the predictable disease
Data set
Attribute measurement using information gain, ratio gain and index gini and pruning of tree
Tree pruning
Result of Classifier

Figure 2 The Naive Bayes method was applied to the patient data.

- **Support Vector**

As a result of the fact that Support Vector Machines have proven to be quite effective in a wide variety of pattern categorization tasks, they have recently attracted a significant amount of interest. The Support Vector Machine, or SVM, is a form of supervised machine learning. The SVM algorithm accurately predicts the occurrence of disease and optimally classifies the classes by constructing the margin among two data clusters. This is accomplished by projecting the disease predicting characteristics in a multidimensional hyper plane. This approach is able to reach great precision because it makes use of kernels, which are nonlinear features. It has been established that the support vector approach (SVM), which has excellent generalisation efficiency, is advantageous in the management of classification responsibilities. The method works on lowering the upper boundary of the generalisation error by using a model that is geared toward reducing the amount of risk that the structure poses. Training a support vector machine is equivalent to finding a solution to a quadratic programme problem that is linearly constrained. The method is frequently applied in the field of medical diagnosis. In the SVM technique, the two variables that have the potential to regulate generalisation are the training mistake and the capacity of the assessed learning machine. Both of these variables are considered to be capacities. The training error rate can be adjusted by making changes to the attributes that are contained inside the classifiers. The precision of the data mining approach varies between the practise sets and the test sets according to the characteristics of the information sets as well as the size of the data sets. When it comes to healthcare data sets, highly unbalanced data sets are one of the most common aspects. This means that the majority classifier and the minority classifier are not in a balanced state, which leads to an inaccurate forecast when the classifiers are run. In addition to these characteristics, information sets for healthcare often contain values that are missing. Because there is typically such a little amount of information that can be accessed, the sample size of the information is frequently considered to be another attribute. There is not an effective method of information mining that can solve all of these issues at once.

- **Neural Network**

It is commonly known that neural networks can generate results in practical applications that are of an incredibly exact nature. The neural network was trained with the database through the utilisation of the feed-forward neural network model, variable learning speed, and the momentum learning algorithm back-propagation. Following is an explanation of the construction of the model: it starts with the input of clinical information and then moves on to the development of an ANN algorithm. Following the completion of the training model, it is able to generate prediction results.

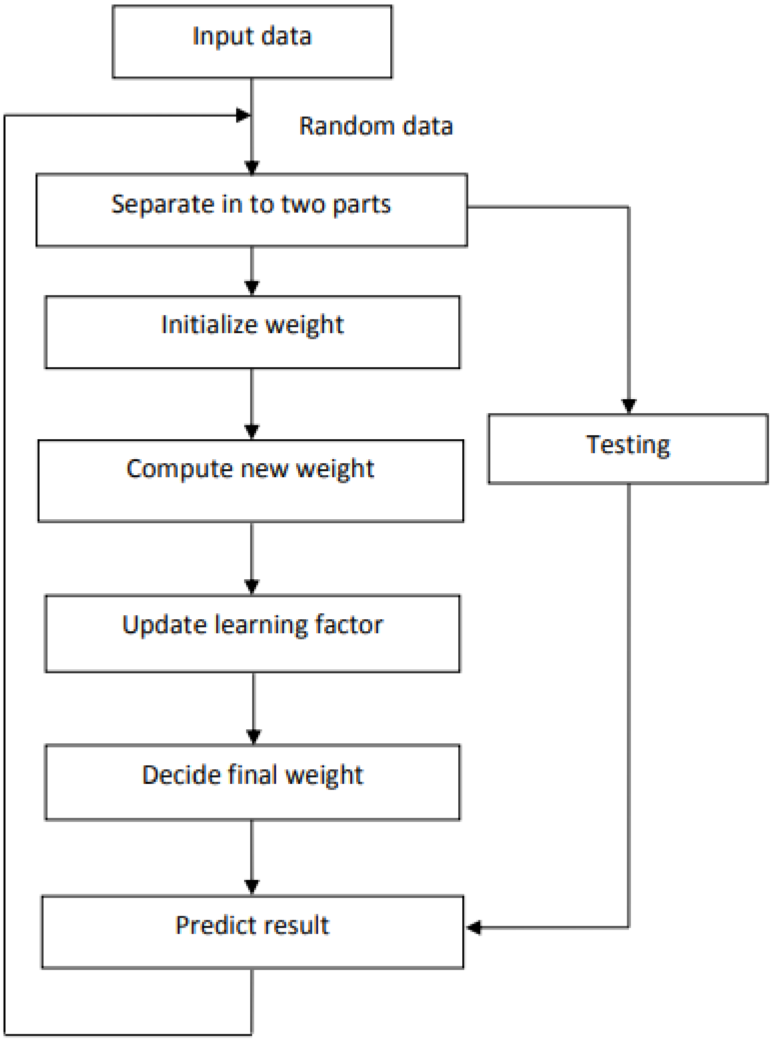

Figure 3 Implementation of neural network algorithm on clinical data.

The random division of clinical data into two equal halves is where the computational stages of an algorithm for a neural network get started. The one that is being used for testing is different from the one that is being used for practise. Randomization is used to determine the initial weight that is given to each function. The mistakes that were calculated are employed in order to modify the weight of all of the qualities. When the mistakes satisfy the termination criteria, the final weight of each feature is determined. The procedure is carried out on a number of different occasions. After constructing the training models, we are able to derive the performance outcomes from the test information and calculate them. Figure 3 depicts the execution of the neural network method with regard to clinical information.

- **Hybrid**

One of the most significant challenges that the healthcare industry is currently experiencing is disease prediction. In order to diagnose diseases, scientists employ a wide variety of information mining approaches. This is motivated by the rising death rate of patients worldwide. Each approach has a set of advantages as well as some drawbacks. Every algorithm that is utilised by every approach incorporates several characteristics that are

ISSN:2320-3714
Volume:3 Issue:3
July 2021
Impact Factor: 5.7
Subject Computer Science

helpful in the process of diagnosing the illness. The term "hybridization" will be used to refer to the output mixture here. In the process of disease diagnosis, the application of hybrid data mining approaches can produce some encouraging outcomes. The methodology behind the hybrid data mining technique is illustrated in figure 4

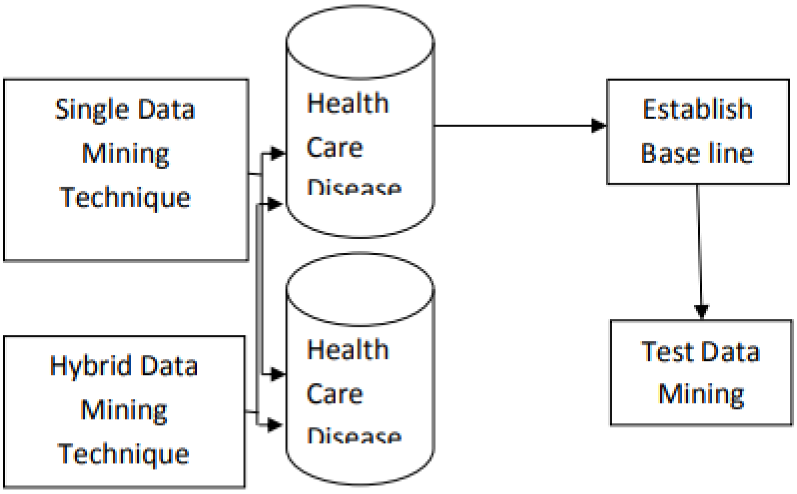

Figure 4 the Approach of hybrid data mining technique.

5. CONCLUSION

In the field of medicine, everyone is concerned with having a high level of accuracy, and recent studies have shown that the desired optimal level of accuracy can be obtained with the application of data mining techniques. The process of carcinogenesis presents scientists with a significant challenge, and there are few techniques available to them for controlling it. The specialised DM tools are now capable of extracting one-of-a-kind patterns and analyses from the data stored in the EDW. This is accomplished by utilising the DM tools along with the data and certain models that are constructed based on the kind(s) of DM jobs.

REFERENCES

1. Hansapani Rodrigo (2021)," Exploratory Data Mining Techniques (Decision Tree Models) for Examining the Impact of Internet-Based Cognitive Behavioral Therapy for Tinnitus: Machine Learning Approach",

2. AntonellaGuzzo (2021)," Process mining applications in the healthcare domain: A comprehensive review",

3. V Klochko et al (2020),"Data mining of the healthcare system based on the machine learning model developed in the Microsoft azure machine learning studio",

4. Adam Bohr et al (2020)," The rise of artificial intelligence in healthcare applications", Artificial

5. Suneeta S. Raykar (2020),” Cognitive Analysis of Data Mining Tools Application in Health Care Services”,

6. S. Dhanalakshmi (2020),” Survey on Information Mining Procedures Utilized in Healthcare Services”

7. Adam Bohr (2020),” The rise of artificial intelligence in healthcare applications”,

8. Sayantan Khanra (2020),” Big data analytics in healthcare: a systematic literature review”,

9. Hui Yang (2020),” The Use of Data Mining Methods for the Prediction of Dementia: Evidence from the English Longitudinal Study of Aging”,

10. Dharmpal Singh (2019),” Cognitive Social Mining Analysis Using Data Mining Techniques”,

11. S. Aarathi (2019),” Impact of Healthcare Predictions with Big Data Analytics and Cognitive Computing Techniques”, International Journal of Recent Technology and Engineering

Author’s Declaration

I as an author of the above research paper/article, hereby, declare that the content of this paper is prepared by me and if any person havingcopyright issue or patent or anything otherwise related to the content, I shall always be legally responsible for any issue. For the reason of invisibility of my research paper on the website/amendments /updates, I have resubmitted my paper for publication on the same date. If any data or information given by me is not correct I shall always be legally responsible. With my whole responsibility legally and formally I have intimated the publisher (Publisher) that my paper has been checked by my guide (ifany) or expert to make it sure that paper is technically right and there is no unaccepted plagiarism and the entire content is genuinely mine. If any issue arise related to Plagiarism / Guide Name / Educational Qualification / Designation/Address of my university/college/institution/ Structure or Formatting/ Resubmission / Submission /Copyright / Patent/ Submission for any higher degree or Job/ Primary Data/ Secondary Data Issues, I will be solely/entirely responsible for any legal issues. I have been informedthat the most of the data from the website is invisible or shuffled or vanished from the data base due to some technical fault or hacking and therefore the process of resubmission is there for the scholars/students who finds trouble in getting their paper on the website. At the time of resubmission of my paper I take all the legal and formal responsibilities, If I hide or do not submitthe copy of my original documents (Aadhar/Driving License/Any Identity Proof and Address Proof and Photo) in spite of demand from the publisher then my paper may be rejected or removed from the website anytime and may not be consider for verification. I accept the fact that as the content of this paper andthe resubmission legal responsibilities and reasons areonly mine then the Publisher (Airo International Journal/Airo National Research Journal) is never responsible. I also declare that if publisher finds any complication or error or anything hidden or implemented otherwise, my paper may be removed from the website or the watermark of remark/actuality may be mentioned on my paper. Even if anything is found illegal publisher may also take legal action against me

Swapna Bhavsar
Dr. Ashish Chaurasia