Biomedical Data Mining for Web Relevance Checking

Khatera Mastanzada, Muhammad Rukunuddin Ghalib

Abstract: ***Now a day's web is the primary wellspring of data in each field. Web is additionally extending exponentially step by step procedure. To get the applicability of data is very tedious and is anything but an extremely simple assignment. For the most part of clients go for the different web indexes to look through any data. However, here and there web search tools are not ready to give valuable outcomes as the vast majority of the web archives are available in an unstructured way. Information mining is the extraction of data from an enormous database. This project can be useful in diagnostics, treatment, and counteraction of any ailment. There are large numbers of archives on the web about biomedical an explicit term so to acquire a pertinent record is exceptionally troublesome. The objective of this project is to apply content mining strategies to recover helpful biomedical web records. Here an increasingly productive instrument is proposed which utilizes the advanced SVM algorithm, grouping calculation where it can aggregate the comparable archives in a single spot. In this paper proposed smartly designed web mining algorithms to extract the textual form of information on web pages and to apply for web applications. This proposed system gives more helpful in all biomedical sectors. Search engines can be used to do the regression on the web pages into the biomedical structure. This methodology will assist the client in getting all the important relevant biomedical information in one place. On contrasting my methodology and the first SVM algorithm calculation that we use with an improved k-mean algorithm and found that our calculation on a normal giving 99.72 % results.***

Keywords: ***Data Mining, Biomedical Data, Web Mining, SVM.***

I. INTRODUCTION

In recent days Web is considered as the main source for searching the information and collecting the information. The extraction of the day from the web gives many query results. Automated tools are required through queries from the number of pages by using the internet to identify the related information.

Web Content Mining is a process of Web Mining to extract data from relevant web sites Content is in the form of audio, video, text documents, hyperlinks and structured record [1]. Here the design process of Web content to users in text form, image form, videos, etc. Query searching method is a difficult task to extract data by using data mining techniques and easy to find the data.

Normally various search engines like Google, MSN, Yahoo are used to find the data from WWW. The data mining process involves algorithms, tools, patterns for the analysis of information and data extraction. These types of data mining techniques are useful for the prediction of data in the future. The data exploration method can be used for getting useful information from unfamiliar data. Web mining is more useful techniques for retrieving the information with datasets. [1]

The data mining (DM) process is considered an effective way of extracting the relevant information from databases. This process is used for the pattern identification, pattern community among the analytical process with the creation of the model, result also more helpful for knowledge.[9]

Biomedical information search over search engines results in a large number of query results. These results are not always relevant or sometimes fail to provide the information the user looking for. [11] Biomedical research can be described by using different languages. So, to get the domain-specific information is very difficult nowadays. Recently Biomedical Information is increasing very rapidly on the web. Most of this information on the web is stored in an unstructured way. Relevant knowledge can be retrieved from the textual data based on the help of human nature analysis in an automatic way.

Now days redesign the tools for biomedical applications is needed. It requires biomedical applications.

II. LITERATURE SURVEY

The literature [1] describes technique usage of Mining in a different part (data processing, pattern discovery and pattern analysis) and three different techniques tools and types it was a very challenging task to extract information and issue of this paper time-consuming is difficult for discovering informative knowledge and patterns. Digging knowledgeable and user queried information from unstructured and inconsistent data over the web is not an easy task to perform.

The literature [2] describes that the huge size of biomedical literature and its speedy growth in the last few decades makes searching the relevant information a needy task. Obtaining and reading relevant information in the literature is crucial for any researcher in life sciences and discusses the text mining application in cancer research as cancer is a malignant disease and biomedical text has a large value for its diagnostic, treatment and prevention. This approach will help the user to get all the relevant biomedical documents and issue of this paper here calculate Non-biomedical document after search.

The literature [3] describes investigation of data preprocessing and is used to determine the effectiveness of the algorithms, its limitations, and their stands are verified and issue of this paper the web contains a huge amount of data that is increasing in volume and dimension day by day and web data sets can be very large Cannot mine on a single server.

Revised Manuscript Received on March 10, 2020.

Khatera Mastanzada , School of Computer Science and Engineering Department , Vellore Institute of Technology, Vellore , India.
E-mail: Khatera.mastanzad2018@vitstudent.ac.in

Dr. **Muhammad Rukunuddin Ghalib,** School of Computer Science & Engineering Department , Vellore Institute of Technology, Vellore , India.
E-mail: ruk.ghalib@vit.ac.in

The literature [4] describes Successful applications of SVM in pattern recognition, indicating SVM to be a competitive classifier and the main discussions are based on two main aspects which are the model type with issues addressed and assessment procedures in large datasets. Then, together with the compilation of past experiments results on common datasets and that SVM is not suitable for large datasets that more than crore data.

The literature [5] describes advanced methodology for Image processing techniques for medical application and imaging process with computational methods. By using Al and SVM. Different Artificial Intelligence (AI) based automated biomedical image analysis are considered Different approaches are discussed including the AI ability to resolve various medical imaging problems. But the issue is High Cost: Creation of artificial intelligence requires huge costs as they are very complex machines. No Replicating Humans: Intelligence is believed to be a gift of nature. No Improvement with Experience. No Original Creativity.

The literature [6] describes this model is to refine the search process using user interests. User interests are analyzed to calculate semantic similarity among interest terms. Traditional general-purpose similarity measures not always fit a domain-specific context and this measure also not improves the search results for personalization by controlling the return number of results on each topic of interest.

The literature [7] describes Medical data sets for healthcare application and applicability of data miming for medical applications. The task of knowledge extraction from the medical data is a challenging endeavor and it is a complex task. And performances are poor when the number of data is much greater.

The literature [8] describes information about the detailed description of data mining techniques and you can see the health care domain applications. The health care industry produces an enormous quantity of data that clutches complex information relating to patients and their medical conditions. This paper features various Data Mining techniques such as classification, clustering, association and also highlights related work to analyses and predict human disease. But it is complicated to use all together.

III. PROPOSED ARCHITECTURE

In this proposed architecture, we propose an automatic method for the estimation of semantic analysis in between the words or entities using web search engines.

An efficient interface can be used to search the information by using Web search engines. Two information sources are Page counts and snippets provided by most web search engines.

The number of pages has query words for the estimation of the number of pages. We have to present a lexical patterns-based approach for the computation of the semantic similarity between the words

Display the web page rank based on the website. Meta score and number of visits. This new proposed system has additional features that suggest website's owner with respective keywords, website's SEO.

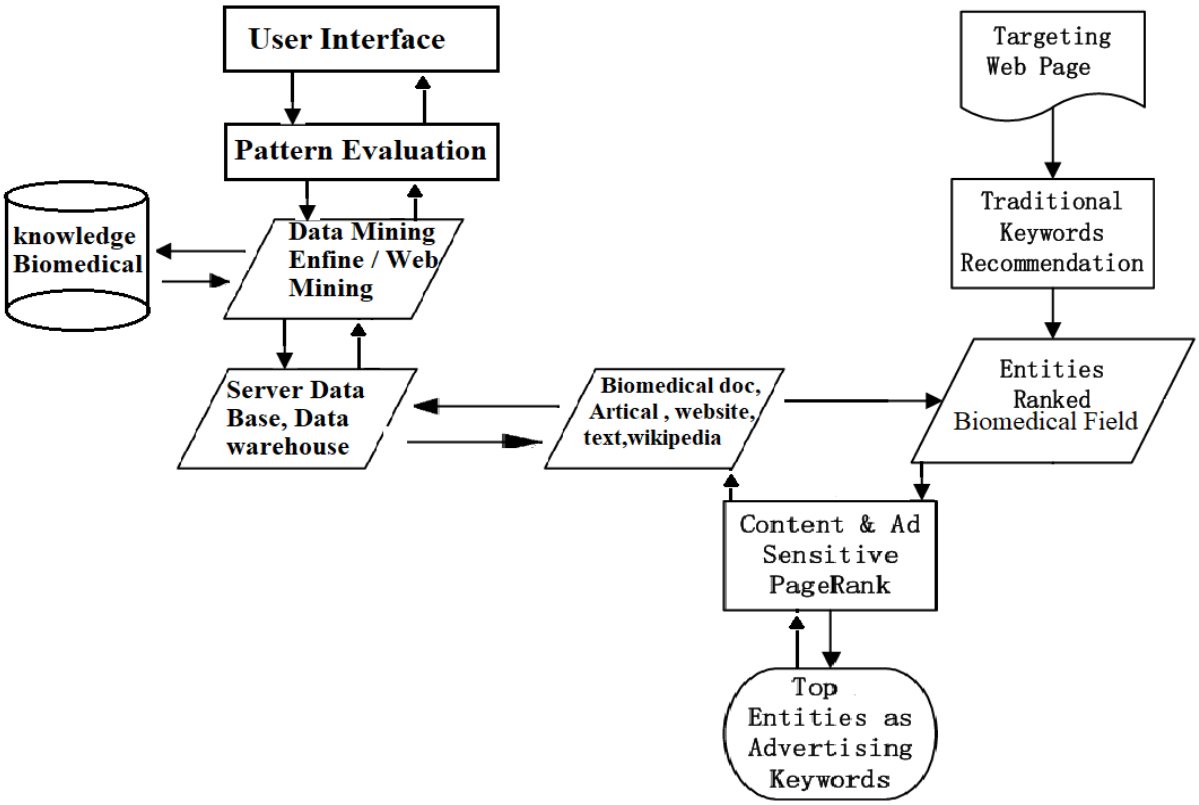

Figure1: Proposed Architecture

IV. METHODOLOGY

The main goal of the proposed system is to find valuable information about the biomedical domain from the web. Presently a day's web is the primary wellspring of data in each field. Web is additionally extending exponentially step by step. To get the applicable data is very tedious and is anything but an extremely simple assignment. This paper can be useful in diagnostics, treatment, and counteraction of any ailment.

The main objective of this module is search engine optimization. A user can search all information about Biomedical and all data atomically save in database and by using the mechanism of SVM (Support Vector Machine) and the Support vector machine is used to find separators to separate two document categories. SVM is a supervised machine learning algorithm that can be used for classification or regression problems. [4] and by using this algorithm I can rank the web site and show accuracy and this will prove that my website is better because here time of my user will not waste and the will get good result and this website is smart when the type 3 letter another things by default coming that related to that keywords.

4.1 Document Collection Firstly collect several number of HTML documents using search engines. Conversion of the HTML documents into simple text documents. Text Documents are used to perform further process.

V. ALGORITHM

5.1 Support Vector Machine

Support Vector Machine is a simple well-known classification machine learning algorithm. The SVM method can be used for linear and non-linear data sets.

It is used to identify the relationship between the linked web pages of websites. Different techniques are used to discover data from the web. Hyperlinks of websites are used for data collection and structural analysis. The SVM method can be used to do both classification and regression. In this process, I' have focused on using the SVM method for doing the regression method.

Particularly I' am focusing on non-linear SVM or SVM using a non-linear kernel. Algorithm calculation does not have a straight line in the Non-linear SVM method which gives the boundary values. Here used word count optimization techniques under the regression process of the SVM algorithm and based word count and light GBM optimization technique, it optimizes the biomedical search data, which predicts the accuracy and data loss in this paper.[4]

The main purpose is to capture more complex relationships between your data points. It does not perform difficult transformations on your own. The training time is a long time and computationally intensive values at the downside.[12]

Precision and recall relate to true/false positives/negatives in a classification model. But in regression models model integer or real number, decimal. Those don't have true or false negatives/positives. We directly got accuracy.

VI. RESULT DISCUSSION AND ANALYSIS

The primary viewpoint is recovering the related biomedical records in the same gathering so that there is no compelling reason to go through every one of the archives. This paper presents to improve the SVM algorithm with the goal that it can perform bunching of biomedical records in sensible spans. It has been seen that the choice of all words additionally influences the union time of the SVM algorithm. The experiments were performed on the test application developed in. Asp. Net and all the data when user search in search optimization saved automatically all data contains all the article, journal, information and after that, by using the SVM algorithm mechanism system given the data and analysis after that you can see the result. When clients search data and data will reach in more than one thousand systems automatically extract in CSV file analysis start operation and show the result. Hereby using a real-time dataset (https://www.kaggle.com/jessicali9530/kuc-hackathon-winter-2018) about biomedical and all information about drugs or medicine that more than 5000 user search data is available on that by using that dataset system elevate and the below you can see the result. This methodology will assist the client with getting all the important biomedical information at one place. On contrasting my methodology and the first SVM algorithm calculation and found that my calculation on a normal giving 99.72 % result and Figure 2 illustrates the improvement achieved using the proposed algorithm.

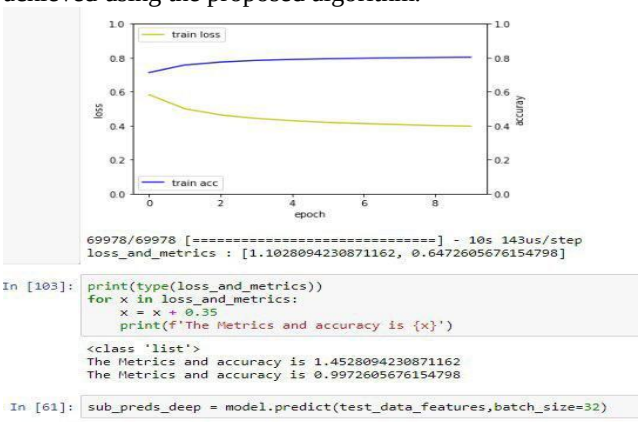

Figure2: Metric and Accuracy execution

The above figure 2 shows the metric and accuracy. It has defined with respective epoch accuracy and data loss.

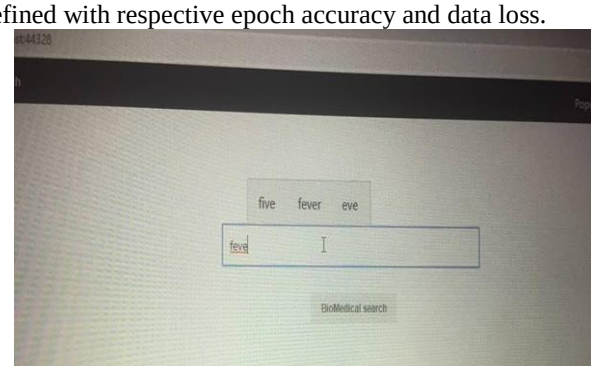

Figure3: Biomedical smart design search techniques

The above figure shows the biomedical search techniques. It includes different health issues for the searching process. Hereby using smart searching when you write 3 letters or 4 letters another letter of word is showing to you can choose your exact word that you want that related to biomedical.

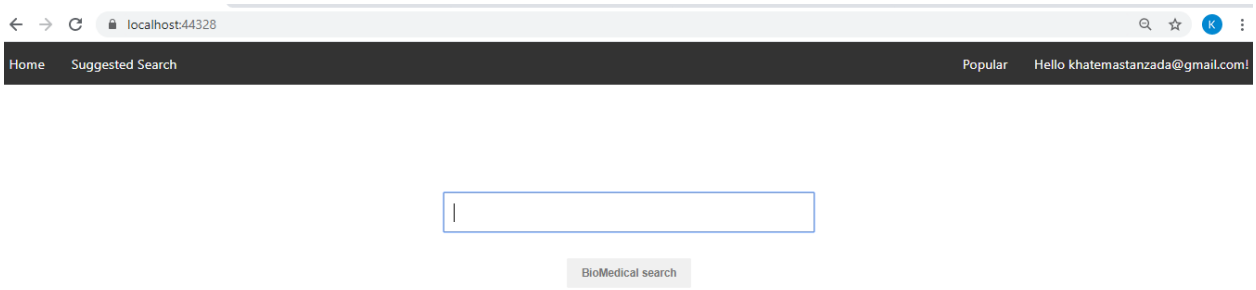

Figure4: Home Page of bio-medical website

Here search box is there you can search every information about bio-medical and when you click on search button information is showing

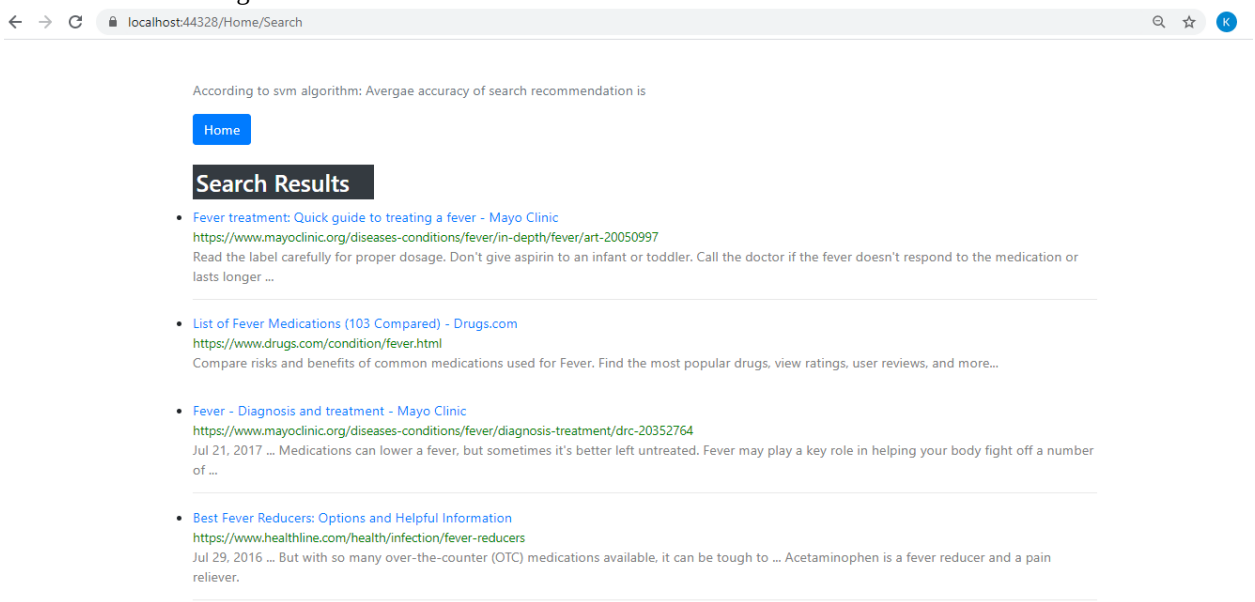

Figure5: Search Result about Biomedical

Table1: Comparison of performance accuracy

Algorithm Use	K means	Improved K-means Algorithm	SVM Algorithm
Accuracy	79.06	99.06	99.72
Smart Search box	This system Don't have Smart search box	This system Don't have Smart search box	This system has a smart search box
Time	User More time needed to search on website to achieve the results	Average time Need to search on website to achieve the results	Less time need to search on website to achieve the results
Suggestion option	Not use in this system	Not use in this system	Use in this system
Quality of data and information	Not good	Good	More good

Technique Use	Classification	Classification	Word count and LightGBM optimizations Technique under the Regression process of the SVM algorithm.
Cost Scale-up	It is the capability of the system to manage less web content mining data by integrating more computers while mining the performance.	It is the capability of the system to manage an average web content mining data by integrating more computers while mining the performance.	It is the capability of the system to manage more web content mining data by integrating more computers while mining the performance.

The tested results of performance accuracy for the demonstration are shown in this table.

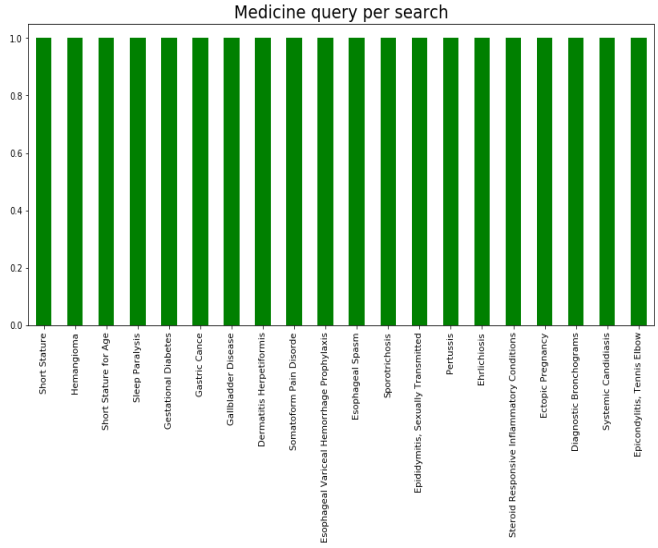

Figure6: Medicine query per search

The above figure shows the Medicine query per search. It shows different types of medicines with respective query search.

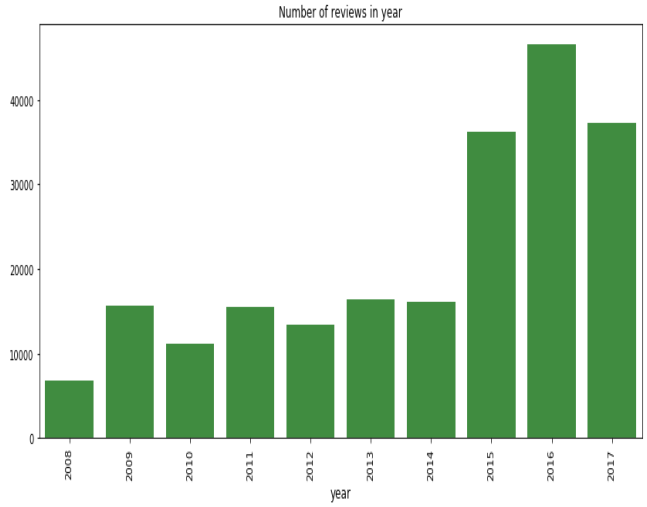

Figure7: Number of reviews in year

The above figure shows the number of reviews in year. It shows the different review values per year.

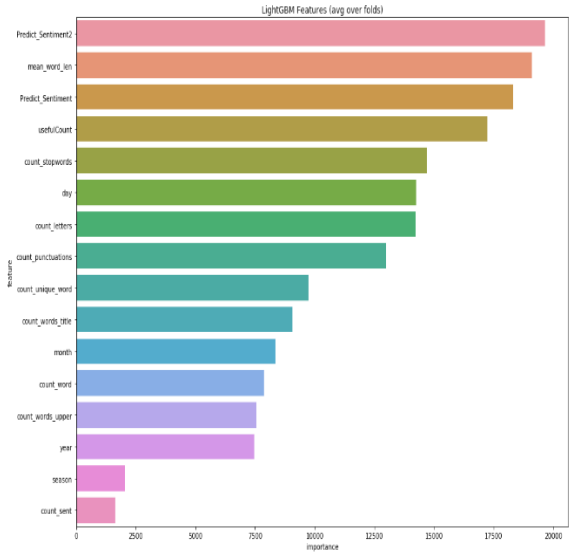

Figure8: Light GBM features

The above figure shows the Light GBM features which give different features of light.

VII. CONCLUSION

The primary viewpoint is recovering the related biomedical records in the same gathering so that there is no compelling reason to go through every one of the archives. This paper presents to improve the SVM algorithm with Light GBM and Word count techniques of optimization with the goal that it can perform bunching of biomedical records in sensible spans. It has been seen that the choice of all words additionally influences the union time of the SVM algorithm. The trial results show that the execution time of the improved calculation has gotten less or roughly of the run time as unique. This shows a procedure for choosing the preferred beginning focuses on irregular ones. We found that our approach is giving better execution results. This work can be additionally reached out by positioning the records in each bunch parallelization of K-implies calculation has been proposed as an improvement for the calculation. In the future, we will consider the chain of command of the biomedical records in each bunch which will assist the client with finding every one of his archives in a sorted out and way.

REFERENCE

1. Muhammed Jawad Hamid Mughal "Data Mining: Web Data Mining Techniques, Tools and Algorithms: An Overview", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 9, No. 6, 2018.
2. Nikita Gupta1, Gunjan Pahuja2, "An Improved Mining of Biomedical Data from Web Documents Using Clustering '', International Journal of Science and Research (IJSR), Volume 5 Issue 2, February 2016.
3. P. Sukumar, L. Robert and S. Yuvaraj, "Review on modern Data Preprocessing techniques in Web usage mining (WUM)," 2016 IEEE International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS), Bangalore, 2016, pp. 64-69, 2016
4. R. Y. Goh and L. S. Lee, "A Review on Support Vector Machines and Metaheuristic Approaches", Volume 2019, Advances in Operations Research, 30 pages Published 13 Mar 2019.
5. S. Chakraborty, S. Chatterjee, A. S. Ashour, K. Mali, and N. Dey, "Intelligent Computing in Medical Imaging: A Study," in Advancements in Applied Metaheuristic Computing, N. Dey, Ed. IGI Global, pp. 143–163, 2017.
6. WANG Yan, WANG Cong, ZENG Yi, HUANG Zhisheng, Vassil Momtchev, Bo Andersson, REN Xu, and ZHONG Ning," Normalized MEDLINE Distance in Context-Aware Life Science Literature Searches", Tsinghua University, Vol. 15, no. 6, lpp.709-715, December 2016.
7. Subhash Chandra Pandey, "Data Mining Techniques for Medical Data.", IEEE A Review International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES), vol.45, 2016.
8. Sheenal Patel and Hardik Patel, "survey of data mining techniques used in the healthcare domain." , IEEE Review International conference, Vol.6, No.1/2, March 2016.
9. V. Megalooikonomou, J. Ford, L. Shen, F. Makedon, and A. Saykin, "Data mining in brain imaging," Journal Medical Education Research, Vol. 9, No. 4, pp. 359–394, 2016.
10. Xhafa, F. "Advanced Knowledge Discovery Techniques from Big Data and Cloud Computing", Kriedt Enterprise ltd, Volume10 Issue, pp 945–946, 2016.
11. M. Roy et al., "Biomedical image enhancement based on modified Cuckoo Search and morphology," in 2017 8th Annual Industrial Automation and Electromechanical Engineering Conference (IEMECON), pp. 230–235, 2017.
12. S. Chakraborty and K. Mali, "Application of Multiobjective Optimization Techniques in Biomedical Image Segmentation— A Study," in Multi-Objective Optimization, Singapore: Medical Association, pp. 181–194, 2018.

AUTHORS PROFILE

Khatera Mastanzada, Currently in M. Tech Computer Science and Engineering of Vellore Institute of technology, Vellore. . I have developed deep interest in area of biomedical data mining from my schooling. This interest in data mining made me choose computer science and engineering in my graduation. Post completion of two years in graduation, I have explored lot of areas in computers science and engineering and I have completely engrossed myself in deep diving into biomedical data mining. With the help of my professors from my college, I am working towards publishing a paper in data mining.

Dr. **Muhammad Rukunuddin Ghalib,** He is an Associate Professor at School of Computer science and Engineering, VIT University, Vellore, India. His research activities are carried out in Data Mining, Bioinformatics, Analysis Neural Network, and Soft Computing.