Comparison of Data Mining Techniques to detect the Brain Tumor

[1]Harmeet Kaur,[2]Jasleen Kaur
[1,2]Department of computer science and engineering
,[1, 2]Khalsa college of Engg. & Technology
Amritsar, Punjab, India,

Abstract :The rapid increase in the Neuro diseases has significantly increased the brain tumor issue to human beings. This work focuses on to classify the brain tumor as yes (tumor infected) or no (not infected from tumor). The detection of brain tumor will be considered as a classification problem and it is evaluated using various state-of-the art algorithms i.e. random forest, Artificial Neural Network (ANN) and Decision-Tree (DT) based mining algorithms. This dissertation has improved the accuracy of mining rate further by ensemble the random forest with Decision-Tree. The simulation result shows that proposed method has better results as compared to the existing technique by using various parameters i.e. true Positive Rate, false positive rate, accuracy, f-measure, precision and recall .

***Index Terms*–Accuracy, Artificial Neural Network (ANN) and Decision-Tree (DT)**

I. INTRODUCTION

Brain tumor may be a cluster of abnormal cells growing within the brain. It's going to occur in somebody at virtually any age. It's going to even amendment from one treatment session to successive however its effects might not be constant for every person. Brain tumors seem at any location, in several image intensities, will have a spread of shapes and sizes. Brain tumors may be malignant or benign. Benign brain tumors have a homogenized structure and don't contain cancer cells. They will be either monitored radio logically or surgically destroyed utterly, and that they rarely grow back. Malignant brain tumors have a heterogeneous structure and contain cancer cells. During this system, we tend to square measure getting to implement a method which may classify tumor and provides additional correct result.

In the data world, data mining is a promising and well known field, attracting a great deal of attention due to the wide availability of data in diverse forms. Moreover, there is an urgent need of turning this data into meaningful and useful information to finally gain knowledge that too at a fast pace. Though, data mining is usually used as a synonym for Knowledge Discovery from Data, but the fact is that it is only an essential part of this radiant KDD process. In simple words,

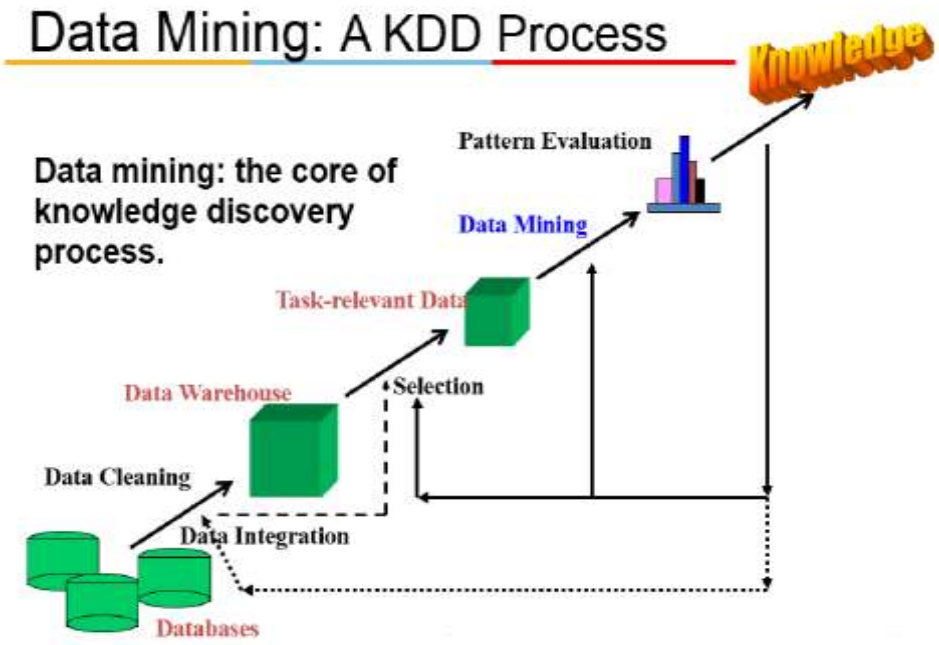

Figure 1.1: Knowledge Discovery from Data (KDD) Process [6]

Data mining is a process where intelligent techniques are applied on preprocessed data (clean, complete, transformed and reduced) for extracting the desired data patterns. It always results in extracting the hidden patterns for decision making. Data mining takes input from the data warehouse or from on-line analytical processing (OLAP) servers. On the contrary, data mining is

considered to be an immense part of the knowledge discovery process, which unfolds into an iterative sequential process depicted in Figure 1.1. Data mining comprises of many intelligent techniques to analyze the given data. Data mining job is not restricted to summarize the data but extract hidden useful patterns for decision making. Based upon the services provided by data systems, the data mining systems can be categorized as:

Machine Learning System: These systems are not capable of handling bulky data. These systems are based on statistical data analysis that follows the rules and procedure of experimental systems.

Database System: These systems (information retrieval) are responsible for retrieving data or information. According to the requirements, the data can be aggregated or generalized. These systems are also capable of answering the query for large databases.

Though data mining can be applied to almost all kinds of data but medical data is one of the most critical data. While working with medical mining, utmost accuracy and fast analysis is required. Impact of data mining is growing day by day on medical data, due to number of reasons like: remarkable increase in medical data size, nearly impossible to manage it manually, more chances of human errors and there adverse effects, heterogeneity of data, endless hidden patterns in given data, incomplete data etc. Data mining could be a great aid to solve these issues to a larger extent.

Reasons for gaining this popularity lies in the fact that data size of medical data is increasing at a very fast pace. Moreover, every stage requires supervision from an expert to derive the best results to help the mankind. It is very tough to perform medical mining due to numerous challenges. Type of medical data is one of the major issues faced by medical miners. Size of medical data is another important challenge to be handled by the data analyzers. Following is the classification of medical data challenges: Domain challenges, General challenges and Collection challenges.

Despite of all these above mentioned challenges, due to sensitivity of the medical data, high risk challenge is another important issue which is faced by medical miners. This is an extremely important issue which is tackled by deriving highly accurate and time effective algorithms for predicting the disease. Not only accuracy but these algorithms must also be versatile and robust in nature. The algorithms must be devised by keeping in mind that these are to be used by number of different users like patients, physicians, nurses and health care system.

II. LITERATURE SURVEY

In medical domain, there is pretty good possibility to work with different medical data sets. The scope of the medical mining largely depends on the type and nature of the medical problem considered. Though, data mining is playing an incredible role in almost every domain but medicine domain is one of the most promising and challenging area. It requires maximum level of accuracy and precision. Moreover, the medical data is a rich source of hidden patterns, whose extraction could be one of the most interesting application areas of data mining. It could be considered as one of the best service to mankind.

Seera, et al. (2014) [2] has proposed a half and half wise framework that comprises of the Fuzzy Min-Max neural system, the Classification and Regression Tree, and the Random Forest model is proposed, and its adequacy as a choice help device for restorative information characterization is analyzed. The cross breed wise framework objective to use the benefits of the part models and, in the meantime, reduces their restrictions. It can gain incrementally from information tests clarify its anticipated yields and accomplish high order exhibitions . To assess the viability of the half breed astute framework, three benchmark medicinal informational indexes, Breast Cancer Wisconsin, Pima Indians Diabetes, and Liver Disorders from the UCI Repository of Machine Learning, are utilized for assessment.

Ali Eman et al. (2015) [4] has displayed a strategy for the order of cerebrum tumors in light of youngsters MRI. The proposed framework comprises of four phases, to be specific, MRI preprocessing stage, Segmentation arrange, Feature extraction, and Classification organize. In the primary stage, the fundamental errand is to dispense with the therapeutic reverberation pictures (MRI) clamor found in pictures because of light reflections or administrator execution which may cause mistakes in the order procedure. The second stage, which is where ROI is separated (tumor locale). In the third stage, the highlights related with MRI pictures utilizing Haar wavelet change (HWT) will be acquired. The highlights of attractive reverberation pictures (MRI) have been diminished utilizing (HWT) to fundamental highlights as it were. Lastly the fourth stages, where new classifier will be introduced lastly the outcome will contrast the proposed classifier and six different classifiers have been utilized. Picture characterization is an essential assignment in the restorative field and PC vision. TANNN will give better outcomes as far as affectability, specificity, exactness and in general running time.

Kumar et al. (2015) [5] has considered that accentuated on the correlation of three bifurcate power based element extraction strategy for the strange examples in cerebrum tumors. Doctor's elucidation of mind tumors may prompt misclassification at some point. There are cerebrum tumors compose Metastatic bronchogenic carcinoma, Astrocytoma, Meningioma, sarcoma. The ascertaining and extricating different power related is utilized character MA TLAB device. GLCM (Gray Level Co-Occurrence) strategy is depict preferred outcomes over alternate techniques J48 calculation furnished in weka with the power based highlights. GLCM is having closest precision to the J48 calculation contrasted with different methods. Our trial result show to be valid that the near examination creates better outcomes with the J48 calculation of WEKA apparatus.

S. Hari Ganesh et al. (2015) [7] presents the latest works completed on EDM and break down their benefits and downsides. This paper additionally features the aggregate consequences of the different information mining practices and methods connected in the overviewed articles, and in this manner proposing the specialists on the future bearings on EDM. Moreover, an examination was likewise led to assess, certain order and bunching calculations to watch the most dependable calculations for future explores.

Ruchika R. Tated et al. (2015) [8] center around the idea of content mining, content mining process, strategies utilized in content mining likewise showing some true utilizations of content mining. Likewise, brief discourse of content mining advantages and impediments has been introduced.

AashimaMalhotra et al. (2016) [10] Introduced novel calculation to order malwares as perfect/ordinary malwares and polymorphic/transformative malwares. The methodology is to produce pydasm report. The guidance sets will be extricated from the

report through content mining and content preprocessing will be improved the situation different procedures like remark expulsion, work extraction and so on.

AndriiShalaginov et al. (2016) [9] investigated use of Neuro-Fuzzy for multinomial classification of malware families and classifications. They gathered a novel dataset comprising of 400k examples for static malware examination. Neuro-Fuzzy performs well considering many-sided quality of the issue and non-linearity of the information.

AashimaMalhotra et al. (2016) [12] broke down different literary works identified with versatile malware discovery. A critical investigation of the terms identified with versatile malware and the strategies utilized for the location of malware is finished. Some proposed strategies and sort of methodologies utilized in those techniques are additionally condensed.

Ramani et al. (2017) [15] has considered on pilgrim look into area because of the difficulties presented by various sorts of pictures and the complexities in accomplishing the precise expectation of variations from the norm nearness. Mind MRI order into ordinary and unusual has gotten expanding consideration on account of the abnormal state of trouble in taking care of those gigantic quantities of pictures. At first, the pictures are pre-handled and the volumetric highlights are separated. At that point, these are nourished into highlight determination procedures viz. Main Component Analysis, Runs, Fisher separating and Relief-highlight choice to decide pertinent highlights.

Mahmud et al. (2018) [16] have been considered distinctive grouping calculations for division show. The basic idea of bunching is to relegate the similitude between the separation, which alludes to the information to quantify the comparability of the measure of the information is requested until the point that every one of the information gathering is done. However, the essential point is to exhibit the examination of the diverse bunching calculations to find which calculation will be most sensible for the clients.

III. METHODOLOGY

In this section we have mentioned the details about the implementation of the proposed protocol and the results found after the implementation.

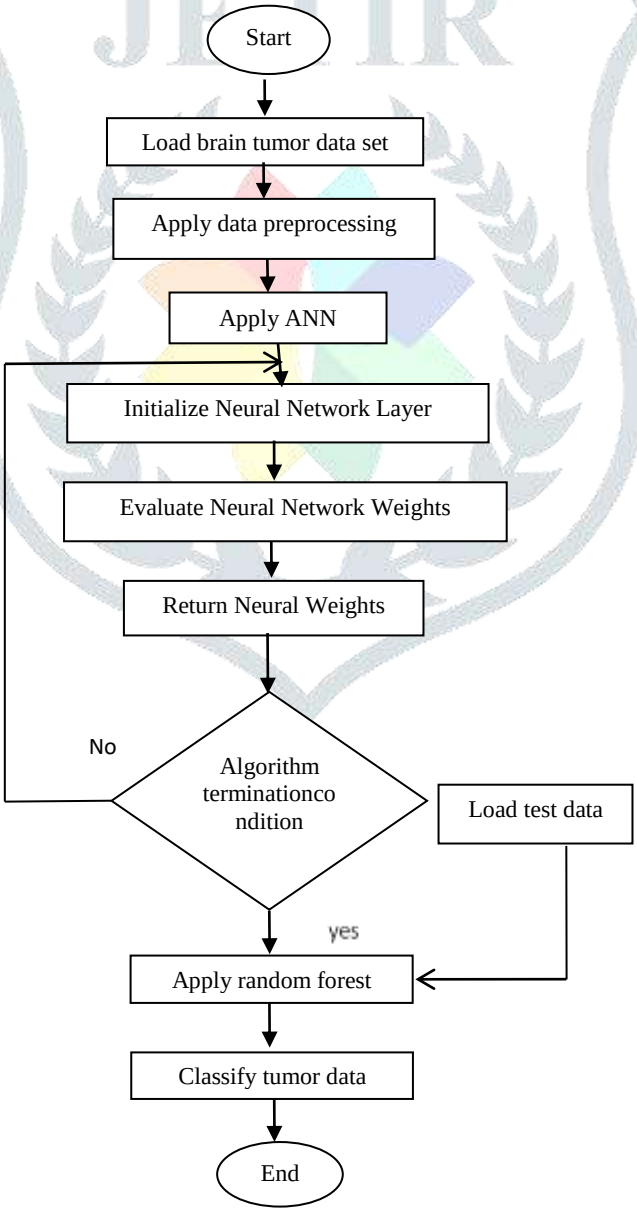

Figure 2.1:Flowchart of the Proposed Technique

IV. RESULTS AND DISCUSSION

This section has shown comparisons between the existing and proposes approach. Tables and Histogram Graphs are produced in order to show the comparison between existing and purposed algorithm. There exist various types of metrics to improve the accuracy rate and the performance of Decision tree based mining algorithms for classification of the scientific applications. Some of them are TP Rate, FP Rate, Precision, Recall, F-Measure, ROC Area which are defined as-

1. TP Rate: TPR refers to True Positive Rate. It is prediction of correctly identified instances. TPR can be expressed by using formula:

$$\text{TP Rate} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Table I: True positive rate evaluation

Iteration	Existing	Proposed
0	0.967	0.99
1	0.971	0.995
2	0.971	0.987
3	0.975	0.99
4	0.967	0.99
5	0.971	0.99
6	0.973	0.988
7	0.974	0.995
8	0.977	0.988
9	0.977	0.988

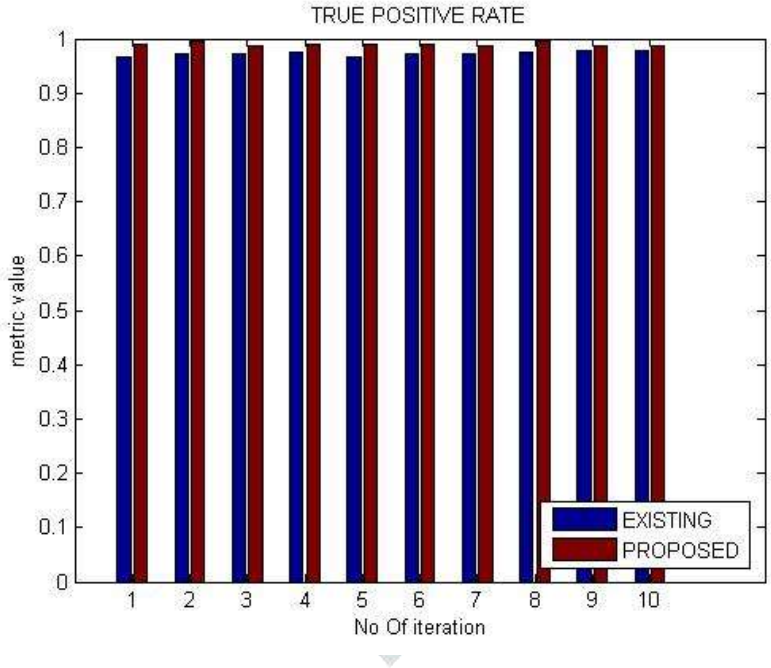

Figure 4.1: True positive rate representation

2. FP Rate: FPR is called False Positive Rate. It is defined as ration of those instances or objects that are incorrectly identified as positive. It is also known as fall-out. FPR can be expressed by using the formula:

$$\text{FP Rate} = \frac{\text{FP}}{\text{FP} + \text{TN}}$$

Table II: False positive rate evaluation

Iteration	Existing	Proposed
0	0.047	`0.014
1	0./043	0.006
2	0.041	0.017
3	0.037	0.014
4	0.047	0.014
5	0.045	0.013
6	0.041	0.015

7	0.038	0.008
8	0.035	0.015
9	0,035	0.015

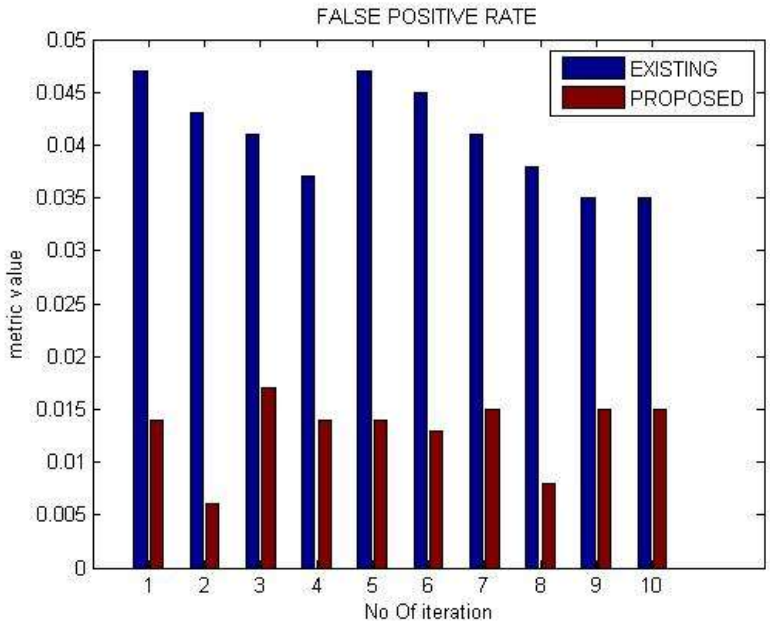

Figure 4.2: False positive raterepresentation

3. Accuracy

The percentage of correctly classified instances over total number of instances.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}$$

Table III: Accuracy evaluation

Iteration	Existing	Proposed
0	96.7448	98.9583
1	97.1354	99.4792
2	97.1354	98.6979
3	97.526	98.9583
4	96.7448	98.9583
5	97.1354	98.9583
6	97.2656	98.8281
7	97.3958	99.4792
8	97.6563	98.8281
9	97.6563	98.8281

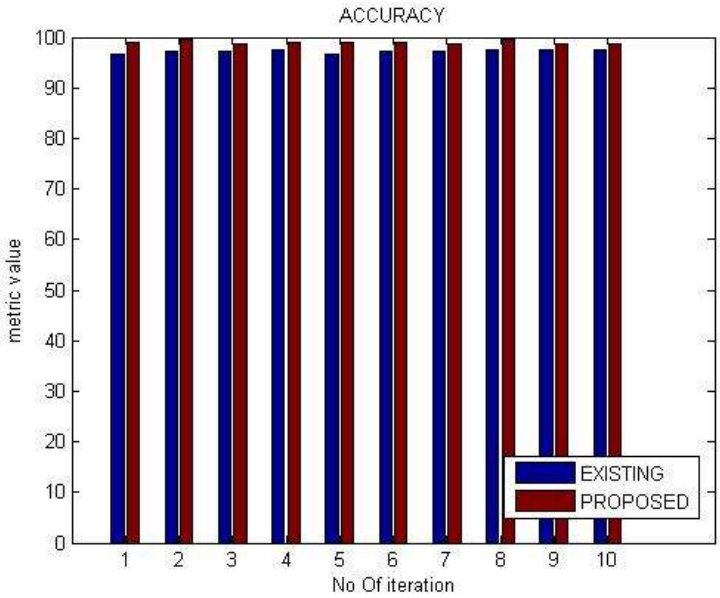

Figure 4.3: Accuracy representation

4. F-Measure: F-Measure is also called F1 score. It computes the mean of precision and recall. Basically, it uses as best and 0 as worst when both precision and recall are used. F-measure can be calculated with using the formula given as:

$$\mathrm{F-Measure} = 2 * \frac{\mathrm{P} * \mathrm{R}}{\mathrm{P} + \mathrm{R}}$$

Table IV: F-Measure evaluation

Iteration	Existing	Proposced
0	0.975	0.99
1	0.971	0.995
2	0.971	0.987
3	0.981	0.99
4	0.967	0.99
5	0.971	0.99
6	0.973	0.988
7	0.974	0.995
8	0.976	0.988
9	0.976	0.988

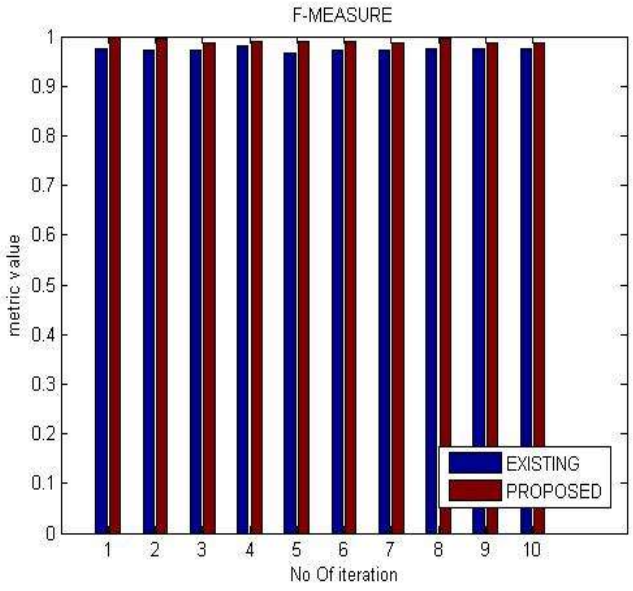

Figure 4.4: F- Measure representation

5. Precision: Precision is defined as measurement of all positive cases that are identified when making calculations. Precision is also known as positive predictive value. Precision can be calculated by using the formula:

$$\text{Precision} = \frac{\text{TP}}{\text{FP} + \text{TP}}$$

Table V: Precision evaluation

Iteration	Existing	Proposed
0	0.967	0.99
1	0.971	0.995
2	0.971	0.987
3	0.975	0.99
4	0.967	0.99
5	0.972	0.99
6	0.973	0.988
7	0.974	0.995
8	0.977	0.988
9	0.977	0.988

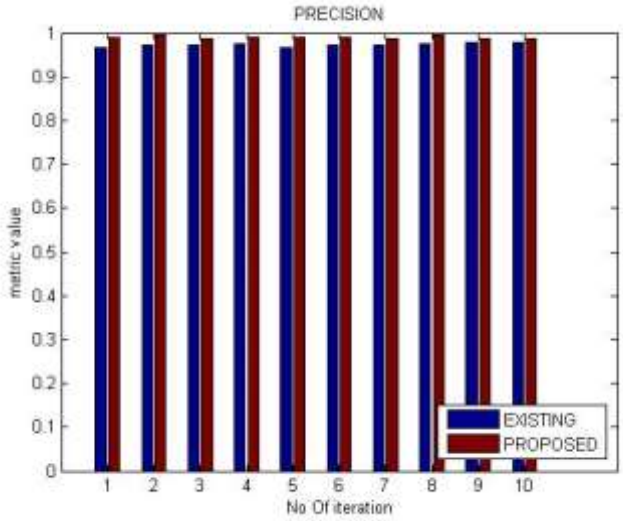

Figure 4.5: Precision Representation

V. CONCLUSION

Data mining plays an important role in analyzing the massive amount of data collected in today's world. However, due to the public's rising awareness of privacy and lack of trust in organizations, suitable Privacy Preserving Data Mining (PPDM) techniques have become vital. A PPDM technique provides individual privacy while allowing useful data mining. It has been observed that existing literature has introduced and evaluates by using various state-of-the art algorithms i.e. random forest, decision trees based mining algorithms but still there are some issues left to enhance the accuracy rate further for classification of the scientific applications.

The proposed technique is designed and implemented in the MATLAB 2013a by using data analysis toolbox. The simulation result shows that proposed method has better results as compared to the existing technique by using various parameters i.e. true Positive Rate, false positive rate, accuracy, f-measure, precision and recall

REFERENCES

[1] Cruz-Monteagudo, Maykel, M. N. D. S. Cordeiro, and Fernanda Borges. "Computational chemistry approach for the early detection of drug-induced idiosyncratic liver toxicity." Journal of computational chemistry 29.4 (2008): 533-549.

[2] Seera, Manjeevan, and CheePeng Lim. "A hybrid intelligent system for medical data classification." Expert Systems with Applications 41.5 (2014): 2239-2249.

[3] Madheswaran, M. and Dhas, A.S., 2014, July. "An adroit naïve Bayesian based sequence mining approach for prediction of MRI brain tumor image". In Computing, Communication and Networking Technologies (ICCCNT), 2014 International Conference on (pp. 1-7). IEEE.

[4] Ruchika R. Tated, Mangesh M. Ghonge (2015) A Survey on Text Mining- techniques and application, In: International Journal of Research in Advent Technology (E-ISSN: 2321-9637) Special Issue, 1st International Conference on Advent Trends in Engineering, Science and Technology "ICATEST 2015", 08 March 2015

[5] Ali, Eman M., Ahmed F. Seddik, and Mohamed H. Haggag. "Using Data Mining Techniques for Children Brain Tumors Classification based on Magnetic Resonance Imaging." International Journal of Computer Applications 131.2 (2015).

[6] Shamsuddin, Rittika, ArvindBalasubramanian, AmitSawant, and BalakrishnanPrabhakaran. "Calculating patient similarity based on respiration induced tumor motion." In Healthcare Informatics (ICHI), 2015 International Conference on, pp. 122-129. IEEE, 2015.

[7] Kumar, B. S., &Selvi, R. A. (2015, March). Feature extraction using image mining techniques to identify brain tumors. In Innovations in Information, Embedded and Communication Systems (ICIIECS), 2015 International Conference on (pp. 1-6). IEEE.

[8] S. Hari Ganesh, A. Joy Christy (2015) "Applications of Educational Data Mining: A Survey. In: IEEE Sponsored 2nd International Conference on Innovations in Information Embedded and Communication Systems" ICIIECS'15

[9] Huda, S., Yearwood, J., Jelinek, H. F., Hassan, M. M., Fortino, G., & Buckland, M. (2016). A hybrid feature selection with ensemble classification for imbalanced healthcare data: A case study for brain tumor diagnosis. IEEE Access, 4, 9145-9154.

[10] AashimaMalhotra, Karan Bajaj (2016) "A Survey on Various Malware Detection Techniques on Mobile Platform, In: International Journal of Computer Applications "(0975 – 8887) Volume 139 – No.5, April 2016

[11] AashimaMalhotra, Karan Bajaj (2016) ".A hybrid pattern based text mining approach for malware detection using DBScan," In: SPECIAL ISSUE REDSET 2016 OF CSIT December 2016, Volume 4, Issue 2, pp 141–149

[12] AndriiShalaginov, Lars StrandeGrini, KatrinFranke (2016) "Understanding Neuro-Fuzzy on a class of multinomial malware detection problems, In: Neural Networks (IJCNN)," 2016 International Joint Conference

[13] Huda, Shamsul, et al. "A hybrid feature selection with ensemble classification for imbalanced healthcare data: A case study for brain tumor diagnosis." IEEE Access 4 (2016): 9145-9154.

[14] Azad, S., Fattah, S. A., &Shahnaz, C. (2017, November). "An automatic scheme for brain tumor region detection from 3D MRI data based on enhanced intensity variation." In Region 10 Conference, TENCON 2017-2017 IEEE (pp. 1-6). IEEE.

[15] Ramani RG, Sivaselvi K. Classification of "Pathological Magnetic Resonance Images of Brain Using Data Mining Techniques. In Recent Trends and Challenges in Computational Models (ICRTCCM)," 2017 Second International Conference on 2017 Feb 3 (pp. 77-82). IEEE.

[16] Mahmud, M.R., Mamun, M.A., Hossain, M.A. and Uddin, M.P., 2018, February. Comparative Analysis of K-Means and Bisecting K-Means Algorithms for Brain Tumor Detection. In 2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2) (pp. 1-4). IEEE.