Journal of Engineering Sciences

Vol 11, Issue 7,July/2020

ISSN NO:0377-9254

CLASSIFICATION OF MEDICAL DATA WITH EFFICIENT DATA CLASSIFICATION METHOD USING MACHINE LEARNING TECHNIQUES

Valeti Anuradha[1], M.praveen kumar[2]

[1] II. Year M.Tech , Department of Computer Science and Engineering, Nalanda Institute of engineering and technology, Sattenapalli, Guntur, Andhra Pradesh, India.

anuradhavaleti14@gmail.com

[2] Associate professor, Department of Computer Science and Engineering, Nalanda Institute of engineering and technology, Sattenapalli, Guntur, Andhra Pradesh, India.

praveenm026@gmail.com

Abstract

Medical data classification is an important data mining issue that has been under discussion for a decade and has attracted a number of researchers around the world. The classification of health data for perfect opinion is a growing field of relevance and investigates the removal of records. The sorting techniques provide the pathologist with invaluable information for the diagnosis and treatment of diseases. The aim of this work was to develop a procedure for the analysis and processing of large amounts of medical data in order to help physicians make better decisions during the diagnosis process and to predict high risk patients for developing a specific disease. The proposed solution is a procedure that combines different techniques that have been used to pre-process, analyze and visualize patient record data in order to predict a high risk of developing a specific disease for the patient. The proposed model is compared with the traditional models and the results show that the proposed model is better than the existing model.

Keywords: Medical Data Classification, Machine Learning Technique, Data Processing, Disease Identification

1. Introduction

The field of research that explores data mining technology to meet the needs of the healthcare industry has a variety of names, such as Healthcare Data Mining , Data Mining in Healthcare [4], Medical Data Mining[8] and Clinical Data Mining[2], depending on the emphasis on data types and applications. Generally, digitally stored health data, possibly collected from multiple information sources and in various forms, is called the Electronic Health Record (EHR). The EHR covers a wide range of information on an individual, such as, but not limited to, person demographics, observations, laboratory tests, diagnostic reports, treatments, therapies, prescriptions and allergies[6].

There is an enormous amount of data generated in the field of medicine[1]. These data, if properly processed, may provide useful information. For example, a medical study could find a link between the successes of the medicine and the height of the patient. The reality may be that, if the data includes information on parameters such as height, body weight, eye color, shoe size and more, some connections may seem to be important only by chance. We can implement large-scale data systems that can analyze DNA in a matter of minutes. When these systems pass through vast amounts of medical data, they find patterns.

Someday, these patterns could lead us to find cures for some of the deadliest diseases. Governments can now analyze data on drugs prescribed to patients by doctors in the public sector. This helps them gain a clear picture of the types of drugs that are prescribed, helping them to understand whether patients are

receiving the most up-to - date medicines. Knowledge discovery in databases (KDD) is the process of automatically generating information formalized in a form that is "understandable" to humans. The KDD Process is depicted in Figure 1.

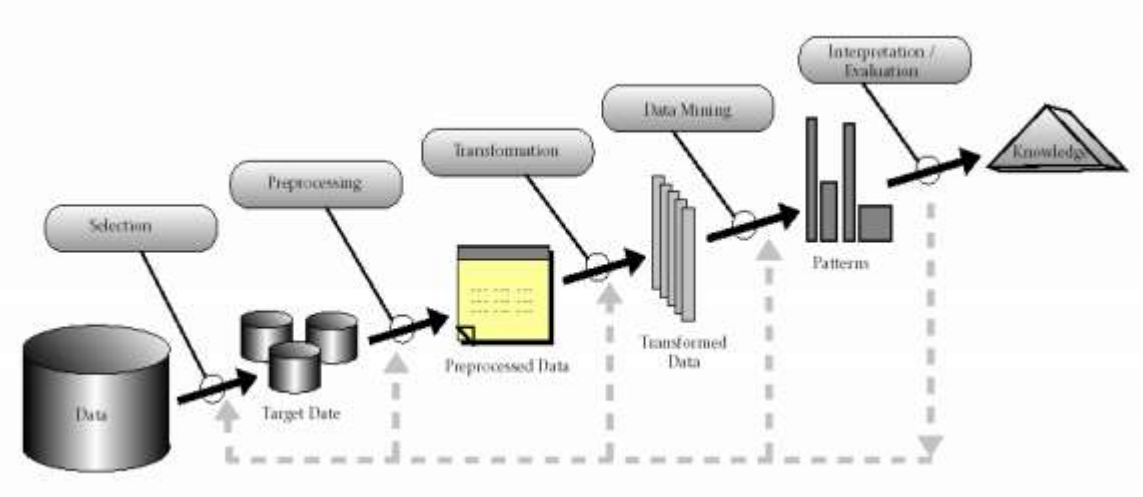

Fig 1: KDD Process

Medical or clinical decision-making support has often been defined as a complex process of collecting, evaluating, analyzing and interpreting medical data for the purpose of formulating one or more decisions, judgments or interventions[4]. It is not a new field in the field of health care or computer science. In fact , people have been managing patient and disease information for decades in order to improve the quality and effectiveness of patient care[3]. Various decision support systems, including medical decision support systems, have been developed in the past since the emergence of knowledge discovery from databases (KDDs)[4]. Data mining techniques have been used to find useful patterns in data, to support medical decision-making, such as diagnostics, choice of treatment options, and prediction of prognosis. When properly applied, it could help the healthcare provider to improve patient care[4].

Due to the high volume and heterogeneity of medical data, it is unlikely that any current data mining tool will work directly with raw data[11]. Pre-processing data is the most important step in performing any type of medical data mining tasks. It is often necessary to identify and eliminate noise, or to deal with inconsistencies and incompatibilities that could slow down computational performance. A reduction in dimensionality is crucial if only relevant information is to be extracted, efficiency improved and computational costs reduced. In fact, it is the most frequent method for pre-processing medical data[8].

In the case of medical data, the reduction in dimensionality can typically be achieved in two ways: record sampling, where some records are randomly selected and used for data mining; or feature selection, where only some (relevant or useful) features are selected. Sometimes the features may be larger than the sample size and, in that case, the latter type of dimensional reduction is preferred. In addition, certain features can only be shown to be relevant when combined. In such cases, the extraction feature, or the process of creating a new feature by combining the existing feature, may also be beneficial.

Although many studies have shown the advantage of the selection of features as part of the pre-processing of medical data mining[5], it must not be ruled out that the process of data integration and information extraction in general may lead to errors[2] and must therefore be carried out carefully. The challenge is not only to extract meaningful information from the data, to discover new insights or patterns, but also to make the data meaningful and to make it useful and usable by the end-user[2]. When done appropriately,

selecting the most important features of medical data could also prevent unnecessary testing and treatment, as they could be costly.

Feature selection in medical data mining is still undergoing active research and new methods or combinations of existing methods are constantly being developed. As both automated feature selection and expert opinions have their own advantages and limitations, the focus of this project is to investigate to what extent medical expert opinions can be incorporated into the automated feature selection for the medical predictive model.

The purpose of supervised machine learning is for a human to provide a machine with data that would consistently train the model and make it smarter and more accurate. There are, of course , different categories of machine learning, such as semi-supervised or unsupervised, which are usually applied to different kinds of problems and domains- but this thesis focuses on supervised learning. The query selection process is depicted in Figure 2.

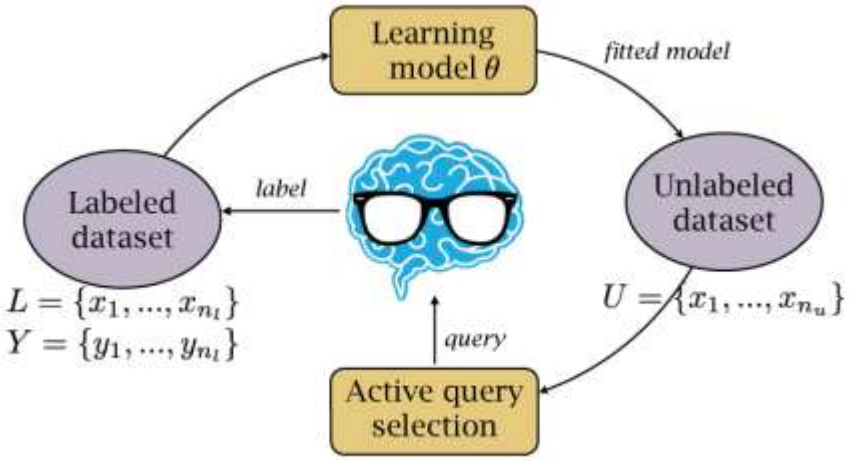

Fig 2: Query Selection Process

Image denoising is a basic operation to improve the analysis of the image. In medical image processing, denosis is a necessary step that is used to detect abnormalities and to highlight some information about the tissues in the image. The denoising process requires prior knowledge of the noise otherwise we can't get the ideal output image.

There are different types of noise in different types of images. The denoising process depends on the problem, the type of image and the noise pattern. While scanning, regardless of the sensor machines, adds a certain amount of noise to the scanned images as well as to the physical phenomena, the scanner properties in the form of noise are added to the image. The presence of noise reduces the visual quality of images and the visibility of low-contrast objects. In general, noise can be introduced by improper machines, acquisition process, transmission and compression. It is an unwanted interference that degrades the visual quality and the information contained in the original image.

2. Literature Survey

Traditional methods (methods used before computers have been introduced to healthcare) use manual analysis to find patterns or extract knowledge from the database. For example , in the case of health care , health organizations (e.g. the Center for Disease Control in the United States) are analyzing trends in diseases and rates of occurrence. This helps health organizations to take future precautions in the

decision-making and planning of health care management.

The traditional method is used to manually analyze data for patterns of knowledge extraction. Take any field, such as banking, mechanics, healthcare, and marketing; data analysts will always work with the data and analyze the final results. The analyst acts as an interface between data and knowledge. Using machine intelligence, we can assist the analyst in producing similar results or data-based knowledge. When we encounter patterns within a database, we report findings (patterns or rules) such as data mining, information retrieval or knowledge extraction, and so on. The term data mining is mostly used by statisticians, data analysts and management information systems (MIS)[7].

Lin, Wenmin et al. [1] proposed a method for the discovery of association rules in medical data to predict heart disease in Andhra Pradesh. This approach was expected to help physicians make accurate decisions. Mortality data from the Registrar General of India show that coronary heart disease (CHD) has been a major cause of death in India. Determining the exact cause of death in rural areas of Andhra Pradesh has shown that CHD causes around 30% of deaths in rural areas.

Chen et al [2] proposed algorithm was evaluated on two manually annotated standard databases such as the CSE and MIT-BIH Arrhythmia databases. This work tells us that the digital band-pass filter was used to reduce the false detection caused by interference in the ECG signal and the further gradient of the signal was used as a QRS-detection feature. They also found that the accuracy of the KNN based classifier was largely dependent on the value of K and the distance metric type. The detection rates for the CSE and MIT-BIH databases were 99.89 percent and 99.81 percent respectively.

Garcia-Hernandez et al. [3] have proposed the choice of a machine learning classification that has been used for the development of a clinical decision support system. Ten sample medical datasets were presented. They suggest the SVM model as the most desirable classification algorithm for developing CDSS. The research was not intended to identify the classification algorithm that was most effective in all medical datasets. More sample datasets have been used to improve the reliability of the model.

Ye and Jun [4] have proposed a growing research on the cardiac disease predictor system, which has become important for the research outcomes categories, and provides readers with an overview of existing cardiac disease prediction techniques in each category. Neural Networks were one of many analytical data mining tools used to make predictions for medical data.

Wójtowicz, Andrzej et al [6] indicated that, in order to evaluate the current general acceptance within the medical community of shaken baby syndrome (SBS), abusive head trauma (AHT) and several alternative explanations for findings commonly seen in abused children. This was a survey of physicians frequently involved in the assessment of injured children in 10 leading children's hospitals. Physicians were asked to estimate the likelihood that subdural hematoma, severe retinal hemorrhage, and coma or death would result from a number of proposed mechanisms.

Thong et al. [8] have demonstrated a method of applying collective intelligence in the field of medical diagnosis by applying consensus methods. They compared the accuracy obtained with that method against the accuracy of the diagnostics achieved through the knowledge of a single expert. The ontological structures of ten diseases have been used. Two knowledge bases were created by putting five diseases into each knowledge base. Two experiments were conducted, one with an empty knowledge base and the other with a populated knowledge base.

3. Proposed Model

Data mining is a key step in the knowledge discovery process in databases where intelligent methods are used to extract patterns [13]. It is the process of analyzing and summarizing data from different perspectives into useful information. The main objective of data mining is to discover new patterns for users and to interpret data patterns in order to provide meaningful and useful information for users. It has been used to find useful patterns to help in the important tasks of medical diagnosis and treatment[1]. Algorithms, when used properly, are capable of improving the quality of prediction,

diagnosis and disease classification[12]. With the data technique, this knowledge can be extracted and accessed by transforming the tasks of the data base by storing and retrieving knowledge and learning [2]. A growing field of application in data mining[7] is the classification of medical data for accurate diagnosis.

The classification of medical data may be used for diagnostic and prognostic purposes. Medical data exhibit unique features, including human noise as well as systematic errors, missing values and even sparseness. The quality of the data has a significant impact on the quality of the mining results. Pre-processing steps are needed to remove or at least address some of the problems associated with medical data[10].

The input data set is selected from the medical database at first. Preprocessing will then be applied to the input medical data set. At the preprocessing stage, we need to extract useful data from the raw medical dataset. After preprocessing, the input dataset will have high dimensional or high characteristics; therefore, the high number of features is a major obstacle to prediction. Therefore, the feature dimension reduction method will be used to reduce the space of the features without losing the accuracy of the prediction. Orthogonal Local Preservation Projection (OLPP) will be used to reduce the dimension of the feature. Once a feature reduction is formed, the prediction will be made on the basis of the optimal classifier. The Group Search Optimizer Algorithm will be used with the Fuzzy neural network in the optimal classifier. The medical data classification process is depicted in Figure 3.

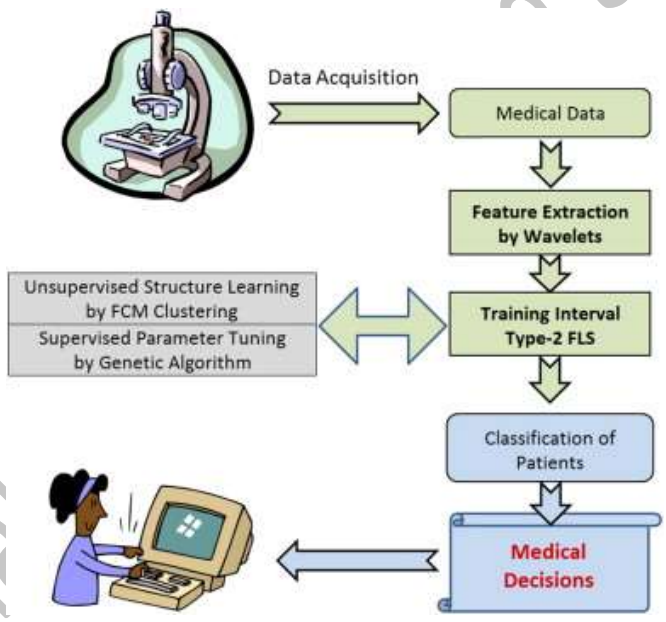

Fig 3: Medical Data Classification Process

The input dataset is given as input at the preprocessing stage. Here, the medical data input has raw data. These raw data are highly susceptible to noise, missing values and inconsistencies. The quality of the raw data affects the results of the method used. In order to improve the quality of medical data and, consequently, the results, raw data are pre-processed in order to improve the efficiency and ease of the mining process. Pre-processing data is one of the most critical steps in the data mining process that deals with the preparation and transformation of the initial data set.

The training set consists of the set of ordered pairs {X, I}, where X={X1, X2, ... Xn} is the input data, and I5-007(1,2, ... m) is the index of one class. The learning process begins by selecting the ordered pair and finding a hyper box for the same class that can be extended (if necessary) to include the input. If it is not possible to find a hyperbox that meets the expansion criteria, a new hyperbox is created and

added to the neural network. The membership function is defined by reference to the minimum and maximum points of the hyperbox. It describes the degree to which the pattern fits into the hyperbox. Hyper boxes range from 0 to 1 in each dimension. The pattern in the hyperbox has a unity membership function. Mathematically, the definition of each hyper box of a fuzzy Hj set is defined by,

$$H_j = \{X, V_{\min_j}, W_{\max_j}, F(X, V_{\min_j}, W_{\max_j})\}$$

Where,

X-Input data,

$V_{minj}(V_{min1}, V_{min2}, \ldots V_{minN})$ is the minimumpoints of H_j

$W_{maxj}(W_{max1}, W_{max2}, \ldots W_{maxN})$ is the maximumpoints of H_j

F (X, V_{mini}, W_{maxi}) is the membership function

Step 1: Initialize the search solution as well as the head angle

- The head angle stated as,

$$\Psi_i^t = \left(\Psi_{i1}^t \ldots\ldots\ldots \Psi_{i(n-1)}^t\right)$$

- The head angle indicated as

$$L_i^t(\Psi_i^t) = \left(l_{i1}^t \ldots\ldots\ldots l_{i(n)}^t\right)$$

- Polar and Cartesian coordinate transformation is depicted as

$$L_{i1}^t = \prod_{p=1}^{n-1} \cos(\Psi_{ip}^t)$$

$$L_{ij}^t = \sin(\Psi_{i(j-1)}^t \prod_{p=j}^{n-1} \cos(\Psi_{ip}^t); \quad \text{Where (j=2}\ldots\text{.n-1)}$$

$$L_{in}^t = \sin(\Psi_{i(n-1)}^t)$$

Step 2: Fitness function is calculated

$$fitness = \min(MSE)$$

Step 3: Find the producer (Z_p) of the group

(i) Scanning operation at zero degree

$$Z_z = Z_p^t + \varepsilon_1 d_{\max} L_p^t(\Psi^t)$$

Where, d_{max} denotes the maximum search distance.

(ii) Scanning operation at the right hand side

$$Z_r = Z_p^t + \varepsilon_1 d_{\max} L_p^t\left(\Psi^t + \varepsilon_2 \Phi_{\max}/2\right)$$

(iii) Scanning operation at the left hand side

$$Z_l = Z_p^t + \varepsilon_1 d_{\max} L_p^t\left(\Psi^t - \varepsilon_2 \Phi_{\max}/2\right)$$

Where, ε_1 points to a normally distributed random number with zero mean and unity standard deviation. And ε_2 stands for a uniformly distributed random sequence that takes value between zero and one. The computation of maximum search distance d_{max}

Maximum search angle Φ_{max}

$$\Phi_{\max} = \frac{\pi}{c^2}$$

The constant c can be stated as:

$$C = round(\sqrt{n+1})$$

Where, n denotes the dimension of the search space.

$$\therefore \Phi_{\max} = \frac{\pi}{n+1}$$

The present best location would take a new best location, if its resource is found as not better than that in the new location. Else, the producer will maintain its location and turn its head according to the head angle direction that is arbitrarily generated.

4. Results

The Proposed model is implemented in ANACONDA SPYDER and the results show that the proposed model is exhibiting better results when compared to traditional models. Most of the datasets (especially those containing categorical / non-sentiment classes) used in the following experiments were balanced and those which were not-have been down-sampled in order to obtain equally class-

balanced datasets. This was required because we use accuracy as our universal metric for evaluating the performance of active learning and, as is known, accuracy does not represent the effectiveness of the model as well as the imbalanced datasets and tasks such as anomaly / outlier detection. In most other tasks, binary classes are usually balanced, so accuracy is a good metric.

Keeping in mind the final objective of the survey, different assessment measurements are used for the effectiveness of our proposed technique. Measurements shall include a collection of considerations, which shall include typical essential assessment strategies. The measurements used here include True Positive, True Negative, False Positive and False Negative, Sensitivity , Specificity and Accuracy.

.

$$Sensitivity = \frac{T(P)}{T(P) + F(N)}$$

$$Specificity = \frac{T(N)}{F(P) + T(N)}$$

$$Accuracy = \frac{T(P) + T(N)}{T(P) + F(N) + F(P) + T(N)}$$

There were four instances of missing characteristics for the status of the lymph node. Dissemination of the class was found to incorporate 151 non-repeat and 47 repetitions. Each record speaks to catch up with information on a single bosom growth case. Alternate features are recorded which condenses each of the 35 characteristics of the dataset. Figure 4 represents dataset quality levels.

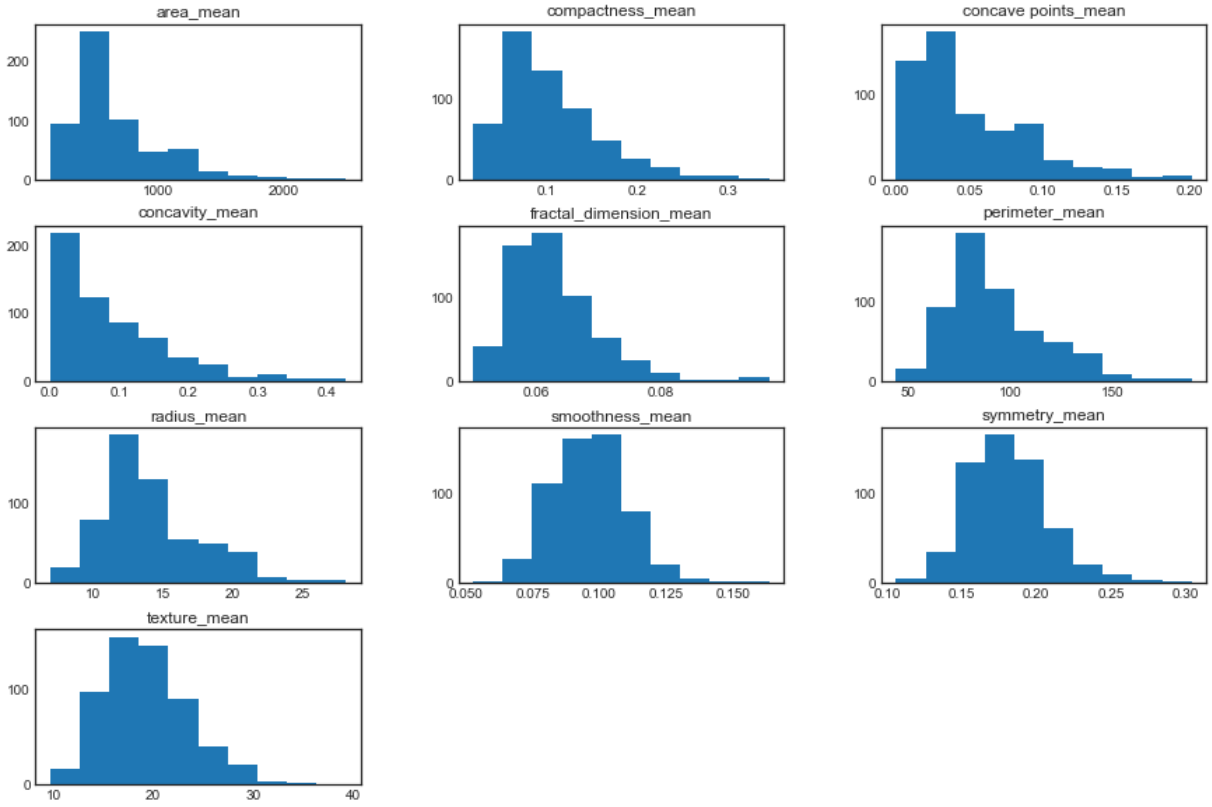

Fig 4: Dataset Quality levels.

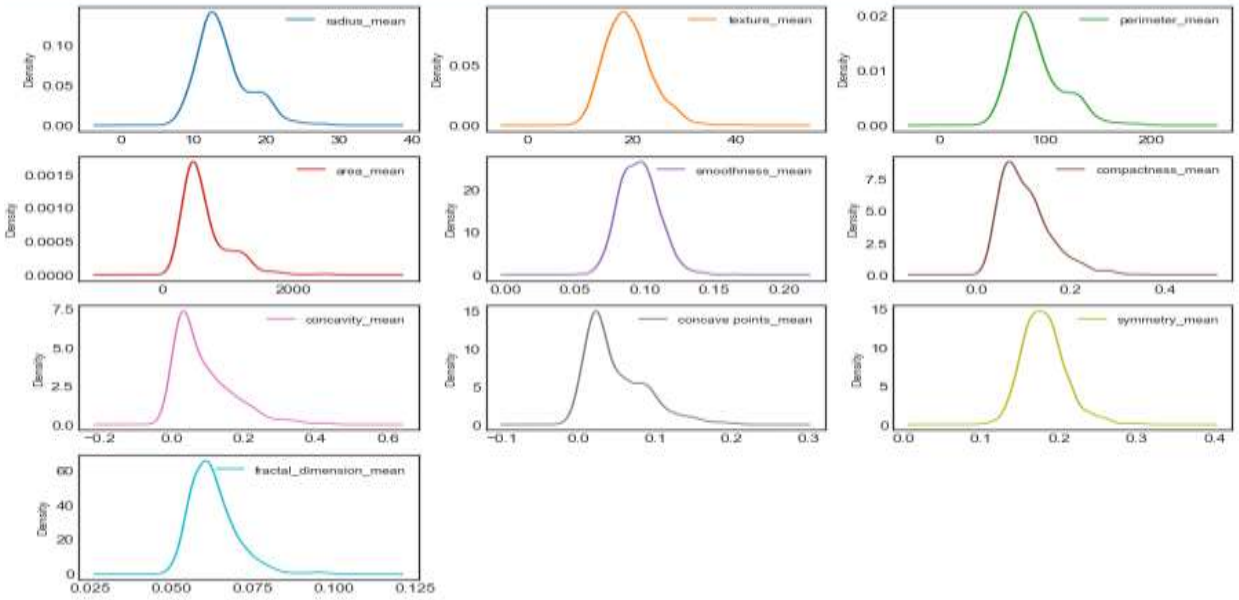

Fig 5: Density levels of Tumor.

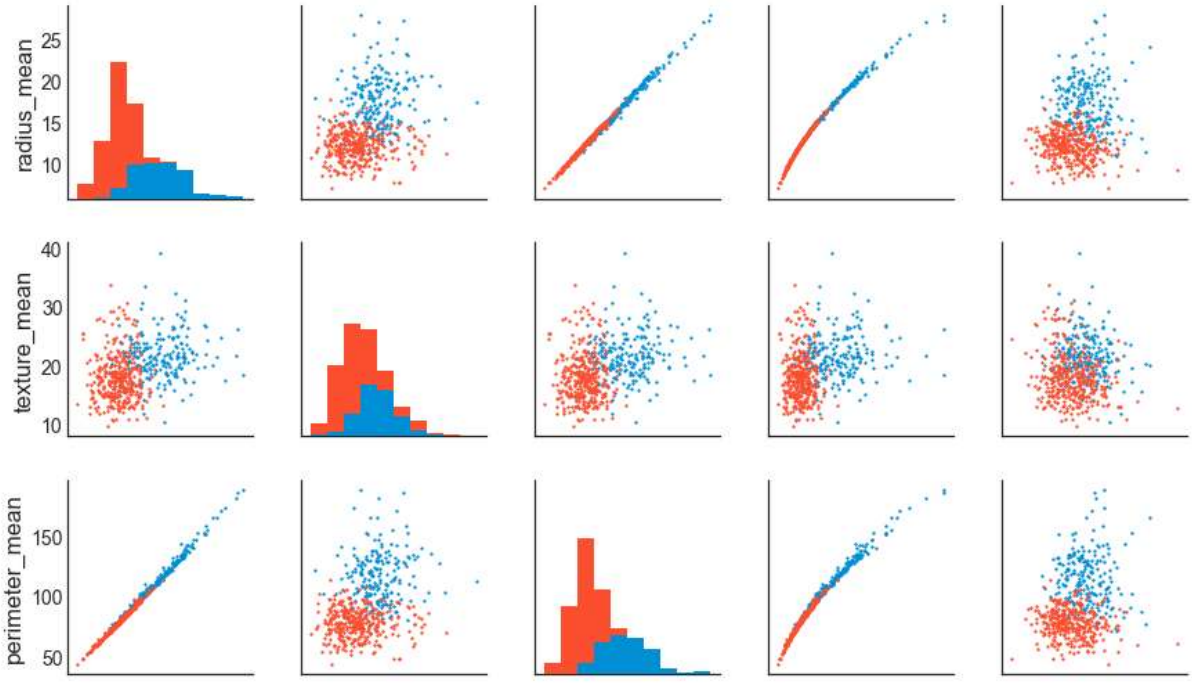

Fig 6: Tumor cluster set.

Feature Selection is the ideal way to find a subset of features from the first feature set. Ant Colony Optimization technique investigates how to locate an ideal subset of features using a few cycles. The primary objective of the proposed technique is to limit the excess between them by selecting a subset

of features. In this technique, the most minimal features of proximity are chosen by each subterranean insect to the features chosen by the past. In this way, if a feature is chosen by the vast majority of ants, this shows that the feature has the most minimal comparability to the alternate features. The feature is best measured by pheromone, and the odds of its determination by alternate ants will be extended in the following cycles. Finally, by making use of the comparability between features, the best features chosen will have high pheromone values. In this way, the proposed technique chooses the best features with the least repetition possible.

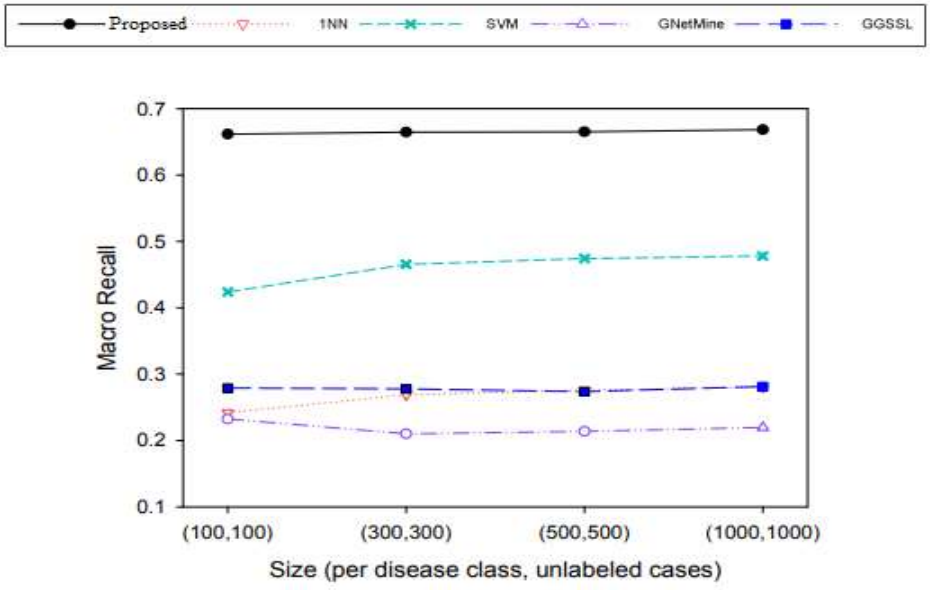

Fig 7 Macro Recall Levels

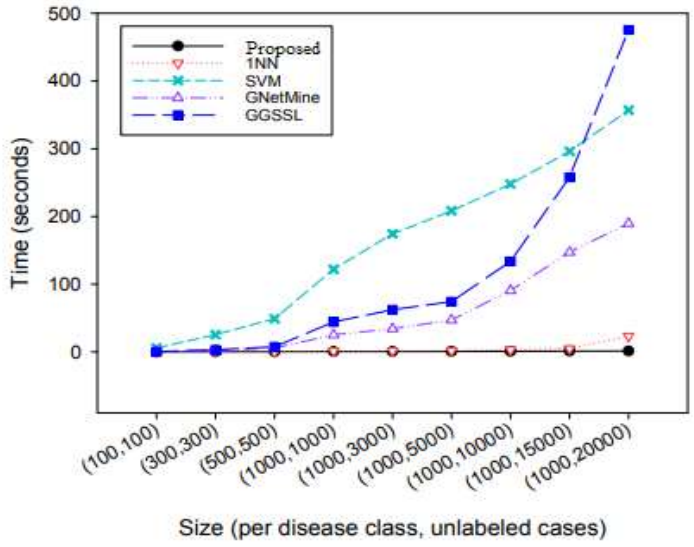

Fig 8: Time for prediction

5. Conclusion

Data on mining health examinations are particularly challenging due to their heterogeneity, inherent noise and, in particular, the large volume of unlabelled data. In this research work, we have recommended a technique for the classification of medical data with the help of an optimal selection of features. Initially, pre-processing is applied to the input data set. The optimal features are then selected from the input dataset using the krill herd algorithm. Next, the selected optimum features are classified using a hybrid adaboost KNN classification algorithm. The results of our proposed method have shown that, compared to the other method, our hybrid classifier achieves better results. Our method achieves the maximum accuracy value for the Hungarian dataset and the Swiss dataset. Both datasets are 97.85 per cent and 98.25 per cent accurate for the classification of medical data.

References

[1]. Lin, Wenmin, Wanchun Dou, Zuojian Zhou and Chang Liu, "A cloud-based framework for Home-diagnosis service over big medical data", Elsevier on Journal of Systems and Software, Vol.102, pp.192-206, 2015.

[2]. Chen, Ling, Xue Li, Yi Yang, Hanna Kurniawati, Quan Z. Sheng, Hsiao-Yun Hu and Nicole Huang, "Personal health indexing based on medical examinations: a data mining approach", Elsevier on Decision Support Systems, Vol.81, pp.54-65, 2016.

[3]. Garcia-Hernandez, Jose Juan, Wilfrido Gomez-Flores and Javier Rubio-Loyola, "Analysis of the impact of digital watermarking on computer-aided diagnosis in medical imaging", Elsevier on Computers in biology and medicine, Vol.68, pp.37-48, 2016.

[4]. Ye and Jun, "Improved cosine similarity measures of simplified neutrosophic sets for medical diagnoses", Elsevier on Artificial intelligence in medicine, Vol.63, No.3, pp.171-179, 2015.

[5]. Thong and Nguyen Tho, "Intuitionistic fuzzy recommender systems: an effective tool for medical diagnosis", Elsevier on Knowledge-Based Systems, Vol.74, pp.133-150, 2015.

[6]. Wójtowicz, Andrzej, Patryk Żywica, Anna Stachowiak and Krzysztof Dyczkowski, "Solving the problem of incomplete data in medical diagnosis via interval modeling", Elsevier on Applied Soft Computing, pp.1-14, 2016.

[7]. Ye, Jun and Jing Fu, "Multi-period medical diagnosis method using a single valued neutrosophic similarity measure based on tangent function", Elsevier on Computer methods and programs in biomedicine, Vol.123, pp.142-149, 2016.

[8]. Thong and Nguyen Tho, "HIFCF: An effective hybrid model between picture fuzzy clustering and intuitionistic fuzzy recommender systems for medical diagnosis", Elsevier on Expert Systems with Applications, Vol.42, No.7, pp.3682-3701, 2015.

[9]. Gorzałczany, Marian B and Filip Rudziński, "Interpretable and accurate medical data classification-a multi-objective genetic-fuzzy optimization approach", Elsevier on Expert Systems with Applications, pp.1-17, 2016.

[10]. Peter Lucas, "Bayesian Networks in Medicine", A Modelbased Approach to Medical Decision Making.

[11]. J.Mago, Parvinder S. sandhu, Neeru Chawla,"Bayesian Network Based Intelligent Pediatric System", World Academy of Science, Engineering and Technology73, pp: 756-760. (2011)

[12]. Carolina A.M. Schurink, Stefan Visscher, Peter J. F. Lucas, Henk J. van Leeuwen, Erik Buskens, ReinierG. Hoff, Andy I.M. Hoepelman, Marc J. M. Bonten, "A Bayesian decision Support System for diagnosing ventilator-associated pneumonia, Springer", Intensive Care Med 33: pp. 1379– 1386, 2007

[13]. S. Devi, A. K. Jagadev, S. Dehuri, R. Mall,"Knowledge Discovery from Bio-medical Data Using a Hybrid PSO/ Bayesian

Classifier TECHNIA – International Journal of Computing Science and Communication Technologies", VOL. 2, NO. 1, July 2009

[14]. Stefan Visscher, Peter J.F. Lucas, Carolina A.M. Schurink, Marc J.M. Bonten, "Modelling treatment effects in a clinical Bayesian network using Boolean threshold functions", Artificial Intelligence in Medicine 46,pp: 251—266, 2009

[15]. Jeffrey Klann, Gunther Schadow, Stephen M. Downs, "A Method to Compute Treatment Suggestions from Local Order Entry Data", AMIA Symposium Proceeding, pp: 387- 391. 2010

[16]. Jyotirmay Gadewadikar, Ognjen Kuljaca, Kwabena Agyepong, Erol Sarigul, Yufeng Zheng, Ping Zhang, "Exploring Bayesian networks for medical decision support in breast cancer detection", African Journal of Mathematics and Computer Science Research Vol. 3(10), pp. 225-231, October 2010

[17]. Stephan Dreiseitla,Lucila Ohno-Machadob, "Logistic regression and artificial neural network classification models: a methodology review",Journal of Biomedical Informatics 35,pp: 352–359, 2002

[18]. Terence D. Sanger, "Optimal Unsupervised Learning in a Single-Layer Linear Feedforward Neural Network", Neural Networks, Vol. 2, pp: 459-473, 1989

[19]. Nikola Masic,G. Pfurtscheller, "Neural network based classification of single-trial EEG data", Elsevier,Artificial Intelligence in Medicine 5 pp: 503-513, 1993

[20]. Yoichi Hayashia, Rudy Setionob, "Combining neural network predictions for medical diagnosis Computers in Biology and Medicine", 32, pp: 237–246, 2002