ISSN 2722-2039

International Journal of Data Science

Volume No. 5
Issue No. 1
January - April 2024

ENRICHED PUBLICATIONS PVT.LTD

JE - 18,Gupta Colony, Khirki Extn,
Malviya Nagar, New Delhi - 110017.
E- Mail: info@enrichedpublication.com
Phone :- +91-8877340707

ISSN 2722-2039

International Journal of Data Science

About the Journal

Data science is the blend of data inferences, algorithm developments, and technology in order to solve analytically complex problems. Data is the core of discussions. Advanced capabilities can be built with it.

The International Journal of Data Science (IJoDS) is an open-access periodical that focuses its discussions on the aspects of data capture, data maintenance, data processing, how to communicate the data, and analyze the data. The journal is an open-access peer-reviewed periodical published biannually. Authors should read the author's guidelines and agree to the copyright and licensing terms prior to submitting the articles.

The Indonesian Society for Knowledge and Human Development (INSIGHT) is a community of scientists. The community office is at the Padang State Polytechnic, West Sumatra, Indonesia. Its members are professionals and researchers in science, engineering, and technology. INSIGHT publishes the International Journal of Advanced Science, Engineering and Information Technology (Scopus Indexed), the International Journal of Data Science, and the International Journal of Halal Research.

ISSN 2722-2039

ISSN 2722-2039

Hairulnizam Mahdin, (Scopus ID: 35759460000) Universiti Tun Hussein Onn Malaysia, Malaysia

Zairi Ismael Rizman, (Scopus ID: 36959761800), Universiti Teknologi MARA (UiTM), Malaysia

ISSN 2722-2039

International Journal of Data Science

(Volume No. 5, Issue No. 01, Januay - April 2024)

Contents

Identification of Gene of Melanoma Skin Cancer Using Clustering Algorithms

Mohanavali Sithambranathan a, Shahreen Kasim a,1,*, Muhammad Zaki Hassan a, Nur Aniq Syafiq Rodzuan a

a Faculty of Computer Science and Technology, Universiti Tun Hussein Onn Malaysia, 86400 Parit Raja, Johor, Malaysia.

ABSTRACT

The Melanoma is the deadliest skin cancer. It can be developed in any parts of the human body. The cancer disease can be cured if it is diagnosed early and proper treatment is taken. In cancer classification, there is a problem in handling the large data of cancer. Large data contains meaningless data and redundant data. Therefore, to overcome the problem, many computer approaches for classification have been proposed in the previous literature. This time, the clustering process for melanoma is conducted using Support Vector Machine and K-Means. Therefore, the purpose of this research is to identify and evaluate the performance of the accuracy of genes that contain melanoma skin cancer using the clustering algorithms.

Keywords:melanoma, skin cancer, identification, gene, clustering algorithms

1. Introduction

Cancer is an abnormal growth of cell (Louise, 2018). There are many types of cancer such as breast cancer, lung cancer, skin cancer and colon cancer. Cancer can be cured if it is diagnosed early and proper treatment is taken. The skin is an organ that separates human body from the environment. Due to that, skin cancer becomes an ordinary type of cancer that affects humans. The number of cases of skin cancer is increasing day by day. Researchers found that the United Kingdom (U.K) is the fast rising country when it comes to skin cancer patients. The two types of skin cancer that exist are melanoma and non-melanoma skin cancer (Wilson, (2012). Melanoma is a type of cancer that develops from the pigment-containing cells known as melanocyte (Talantov, Mazumder and Jack, 2005). Non-melanoma skin cancer refers to all the non-melanoma types of cancer that occur in the skin (Wolters Kluwer, 2019). Skin cancers occurs due exposure to sun and also affected from the genetic problem (Moore, 2001).

In the previous literature, the problem in classifying a cancer is when the gene expression data is used (Lu and Han, 2003). Gene expression data is considered a high-dimensional type of data. Therefore, the analysis of gene expression is difficult to conduct because of the big data that contains noise, redundant data, and unrelated information of features.

In this research, K-Means and Support Vector Machine were used to cluster and classify data and compare the detection accuracy. Clustering involves assigning data points to a cluster where items in the same cluster are the same. Therefore, the clusters are known by some similarity measures for example distance, connectivity and intensity (Narongsak and Anongnart, 2016, Bernhard, 2001).

2. Literature Review

This section discusses the finding of literature reviews related to the research.

2.1. Support Vector Machine

A supervised learning system is the Support Vector Machine (SVM) George et al., 2011. SVM based algorithms used for identification and regression processing to analyze data and understand trends. An

algorithm for SVM learning creates a prototype that assigns new examples to one or the other group, rendering it a non-probabilistic conditional linear classifier.

George et al., 2011 stated that SVM is the best cancer classification system to use. There are several explanations why in cancer classification, SVM has the best performance. Next, SVMs have checked the potential not only to correctly classify organizations into relevant categories, but also to distinguish situations where the evidence does not help understand the classification. SVM has many computational features that make them interesting in the study of gene expression, including their stability in selecting a similarity variable, the lack of solution while dealing with large data sets, the ability to handle wide field spaces, and the ability to identify outliers. To construct a classifier the following formula is used.

$$y\,(x) = \text{sign}\left[\sum_{k=1}^{N} {}_k\, y_k\, \psi\,(x,\, x_k) + b\right]$$

This formula consists of real constant and polynomial SVM degree

2.2. K-Means

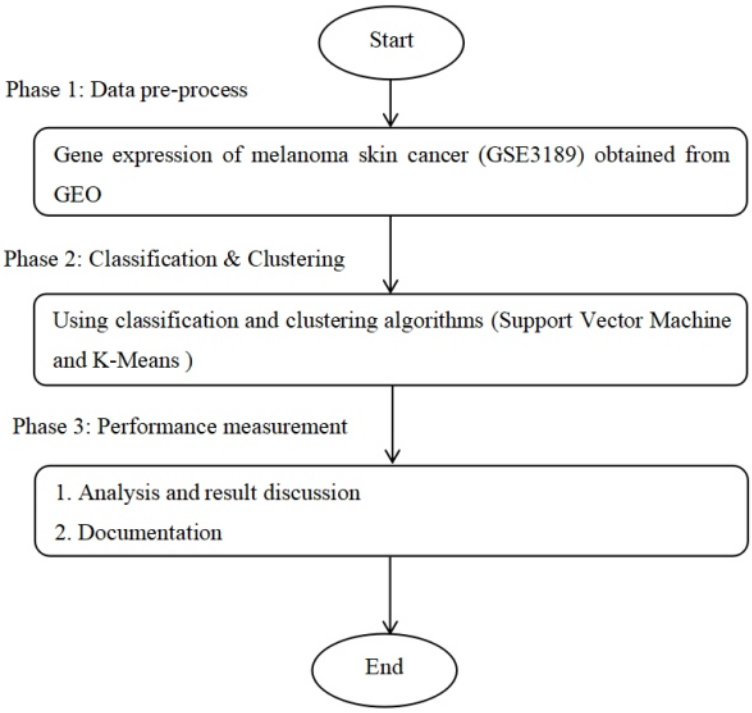
K-means is an incremental clustering method which dynamically integrates one cluster center at a time by way of a deterministic global search process consisting of N (with N being the width of the data set) executions of the k-means algorithm from correct initial positions (Lozano et al., 1999). The K-means method in data mining begins with a first group of randomly selected centroids, which is used as the starting points for each cluster, and then conducts iterative (repetitive) calculations to refine centroid *(ci and cj, ci ≠ cj)* locations. The new means (centroids) of the observations are then calculated in the new clusters.

$$m_i^{(t+1)} = \frac{1}{\left|S_i^{(t)}\right|} \sum_{z_j c S_i^{(t)}} x_j \qquad (2)$$

The calculation shows that the algorithm has converged when the assignments no longer change. The algorithm does not guarantee that the optimum can be found. The algorithm is often presented as assigning objects to the nearest cluster by distance.

3. Methodology

In the first phase which is data pre-processed, the data is collected from GEO database. Melanoma skin cancer gene expression (GSE3189) is obtained from the GEO database. Affymetrix Human Genome U133A Array is the basis of the software used. GSE3189 contains three types of classes namely normal skin, nevi skin and melanoma (Lu and Han, 2003). There are 70 samples in this set of data. 7 are normal skin, 18 are nevi skin and the remaining 45 are melanoma. Next, pre processed data is used in the clustering and classifying process where it uses the Support Vector Machine (SVM) and K-Means. In phase three, the documentation of paper works, soft materials, and coding design used in the research are prepared in the form of paper works. The purpose of documentation is to provide a clear understanding to the readers about the overall flow of the research.



For SVM the first step is to import the dataset into the R environment. Then, the specific code that describes the SVM function to plot the graph is used. The gene expression and selected genes are used as the input and parameter for the SVM. The results will be produced in the format of graph.

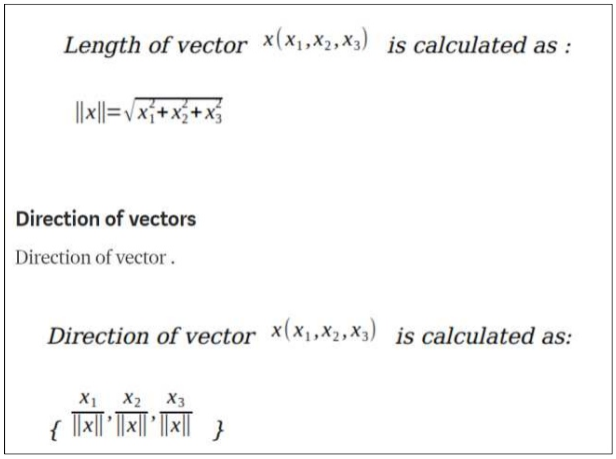


Fig 2. This is a mathematical equation used to calculate the length and direction of vector

As for K-Means the same method is used where first of all the dataset needs to be imported into the R environment. Next, the K-Means function is used to cluster the data of genes into groups. The results will be produced in 2D representation where the genes will be grouped according to the type of skin. The algorithm works as follows:

Step 1: Choose groups in the feature plan randomly.

Step 2: Minimize the distance between the cluster center and the different observations (centroid). It results in groups with observations.

Step 3: Shift the initial centroid to the mean of the coordinates within a group.

Step 4: Minimize the distance according to the new centroids. New boundaries are created. Thus,

observations will move from one group to another

Repeat until no changes are observed in groups.
K-means usually takes the Euclidean distance between two features :

$$distance\ (x, y) = \sum_i^n (x_i - y_i)^2 \quad (3)$$

Different measures are available such as the Manhattan distance or Minlowski distance. It is noted that K-mean returns different groups each time you run the algorithm. We recall that the first initial guesses are random and the distances are computed until the algorithm reaches a homogeneity within groups. This means that k-mean is very sensitive to the first choice, and unless the number of observations and groups is small, it is almost impossible to get the same clustering.

4. Result and Discussion

There are two types of algorithms used to conduct this research which is SVM and K-Means. SVM is the supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. K-means is simple where it groups similar data points together and discover underlying patterns. Figure 3 and 4 show the results for the classifiers' performance.

Fig 3. Result of normal and melanoma for GSE3819/206403 using Support Vector Machine

Figure 3 shows the results of SVM. SVM is a subclass of supervised classifiers that attempt to partition a feature space into two or more groups. The separation boundary is linear, leading to groups that are split up by lines (or planes) in high-dimensional spaces. y as the response variable and other variables serve as the predictors. The data frame will have unpacked the matrix x into 2 columns named x1 and x2. Based on the result, the number of support vectors is 6 and they are the points that are close to the boundary or on the wrong side of the boundary. The support vector 6 is the number of o in the graph. The blue color part indicates the melanoma skin genes while the pink color indicates the normal and nevi skin genes. The points that are close to the boundary are colored blue while the wrong side boundary is pink. The wrong side of the boundary shows that the observed data is sufficiently inconsistent. This shows that the dataset is not grouped properly as the dataset is mixed up as shown in the graph.
Based on figure 4, to perform the analysis, two groups of skin that contain different genes were selected. The data is from the same source and it is tested using k-means algorithm. According to the result the genes are clustered into two groups. It explains the point variability where a centroid is the imaginary or real location representing the center of the cluster. The medium size grouped data is nevi and the largest

grouped data is melanoma skin. A cluster refers to a collection of data points aggregated together because of certain similarities. Therefore, by referring to the result in figure 3, one can conclude that the highest number of centroid in one place is considered as normal cell. According to the datasets there are 45 genes that are melanoma while 18 are nevi. Therefore, the clustering process was carried out to identify groups of similar genes and to group the genes according to the group to which they belong.

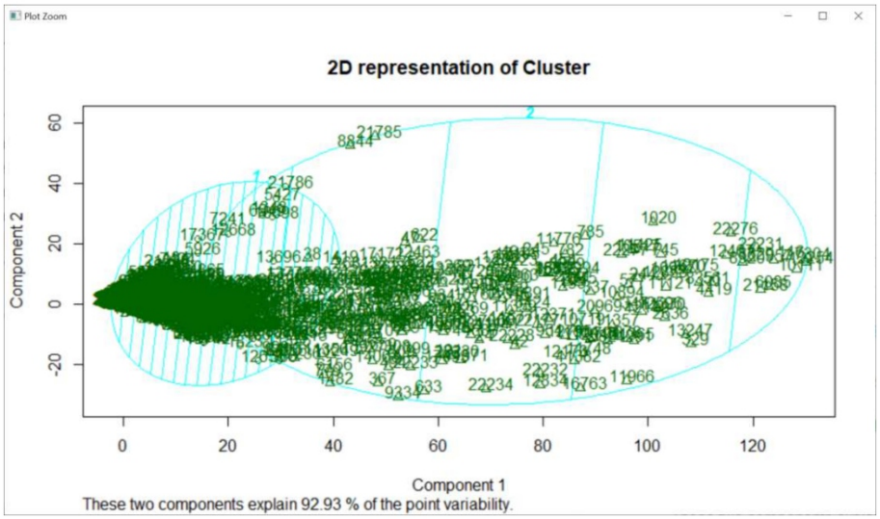


Fig 4. Result of 2D representation for GSE3819 using K-Means

Thus, the overall results obtained from both SVM and K-Means show that the graphs produced by both algorithms are different from one another. According to the result of the SVM graph, there are 2 boundaries that consist of points which are x and o. If the value of X2 is 0.01 then the value below it is considered as the data that is sufficiently inconsistent. As for the K-means graph the points which are in the number form are grouped into two labeled as 1 and 2. The largest grouped data is known as melanoma skin according to the raw dataset calculation.

5. Conclusion

As a conclusion, this research focuses on identifying the result of clustering using K-Means and Support Vector Machine. It is to determine the high accuracy of clustering towards the melanoma data. From the clustering process, the genes will be identified which are most related to the melanoma skin cancer. In the process of classifying the cancer, the data sets of genes expression will be collected. To get the most suitable data for this research, the pre-processed data sets will be done first. The best genes will be obtained through the process of selecting the features using limma package that is used to calculate and normalize the data. Other than that, the package also focuses on differential expressed genes analysis. To find out the accuracy of the selected genes, a classifier will be constructed once the pre-processed data sets are done. Based on the results, K Means is good at grouping dataset compared to SVM. This is because the results obtained from the K-Means are grouped into two according to their criteria but for SVM the results were mixed.

Acknowledgment

I would like to express my special thanks to my supervisor Prof. Madya Dr. Shahreen Binti Kasim for the support and guidance in helping me throughout the course of my research and for simply not giving up on me.

References
[1] Bernhard Scholkopf, Alexander J. Smola. (2001). Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond.
[2] Victo Sudha George, Cyril Raj. Review On Feature Selection Techniques And The Impact Of Svm For Cancer Classification Using Gene Expression Profile. International Journal of Computer Science & Engineering Survey [Internet]. Academy and Industry Research Collaboration Center (AIRCC); 2011 Aug 30;2(3):16–27. Available from: http://dx.doi.org/10.5121/ijcses.2011.2302
[3] Peña J., Lozano J., Larrañaga P. An empirical comparison of four initialization methods for the K Means algorithm. Pattern Recognition Letters [Internet]. Elsevier BV; 1999 Oct;20(10):1027–40. Available from: http://dx.doi.org/10.1016/s0167-8655(99)00069-0
[4] Mulryan C. Understanding cancer: the basics. British Journal of Healthcare Assistants [Internet]. Mark Allen Group; 2010 Jun;4(6):266–9. Available from: http://dx.doi.org/10.12968/bjha.2010.4.6.48484
[5] Lu, Y., & Han, J. (2003). Cancer classification using gene expression data. Information System, 28(4), 243-368.
[6] Hang X. Cancer classification by sparse representation using microarray gene expression data. 2008 IEEE International Conference on Bioinformatics and Biomeidcine Workshops [Internet]. IEEE; 2008 Nov; Available from: http://dx.doi.org/10.1109/bibmw.2008.4686232
[7] Georgia Moore. A course of action on skin cancer. Business and Health 2001;4:40.
[8] Chayangkoon N, Srivihok A. Two Step Clustering Model for K-Means Algorithm. Proceedings of the Fifth International Conference on Network, Communication and Computing - ICNCC '16 [Internet]. ACM Press; 2016; Available from: http://dx.doi.org/10.1145/3033288.3033347
[9] Talantov D. Novel Genes Associated with Malignant Melanoma but not Benign Melanocytic Lesions. Clinical Cancer Research [Internet]. American Association for Cancer Research (AACR); 2005 Oct 15;11(20):7234–42. Available from: http://dx.doi.org/10.1158/1078-0432.ccr-05-0683
[10] Wilson MA, Nathanson KL. Molecular Testing in Melanoma. The Cancer Journal [Internet]. Ovid Technologies (Wolters Kluwer Health); 2012;18(2):117–23. Available from: http://dx.doi.org/10.1097/ppo.0b013e31824f11bf
[11] Corner C, Hoskin P. Skin cancer. Oxford Medicine Online [Internet]. Oxford University Press; 2013 May; Available from: http://dx.doi.org/10.1093/med/9780199696567.003.0018

Evaluate the Performance of SVM Kernel Functions for Multiclass Cancer Classification

Noramalina Mohd Hatta a,1, Zuraini Ali Shah a,2, Shahreen Kasim b,3,*

a Department of Software Engineering, Faculty of Computing, Universiti Teknologi Malaysia, 81310 Johor Bharu, Johor, Malaysia.

b Faculty of Computer Science and Technology, Universiti Tun Hussein Onn Malaysia, 86400 Parit Raja, Johor, Malaysia.

ABSTRACT

Multiclass cancer classification is basically one of the challenging fields in machine learning which a fast growing technology that use human behaviour as examples. Supervised classification such Support Vector Machine (SVM) has been used to classify the dataset on classification by its own function and merely known as kernel function. Kernel function has stated to have a problem especially in selecting their best kernels based on a specific datasets and tasks. Besides, there is an issue stated that the kernels function have a high impossibility to distribute the data in straight line. Here, three basic kernel functions was used and tested with selected dataset and they are linear kernel, polynomial kernel and Radial Basis Function (RBF) kernel function. The three kernels were tested by different dataset to gain the accuracy. For a comparison, this study conducting a test by with and without feature selection in SVM classification kernel function since both tests will give different result and thus give a big meaning to the study.

Keywords: multiclass cancer classification support vector machine kernel functions machine learning SVM classification

1. Introduction

In bioinformatics fields, genes identifications are responsible for classifying existing disease samples of two or more of its variants. As previous study had been done and solved involving supervised learning methods such as k-nearest neighbour (KNN), weighted voting approach, support vector machine (SVM), linear discriminant analysis (LDA), artificial neural networks (ANN) and Random forest. Besides, in cancer classification using microarray data, an increasing number of studies have successfully demonstrated the effectiveness of state-of-the art supervised machine learning methods such as Support Vector Machines (SVMs). SVMs are defined as powerful classification machine learning based on the variety of regularization technique (Niijima and Kuhara, 2005). The SVMs are built to learn a function that generates output based on input and for the next new output can be easily generated since the old function has learned from the previous case.

For gene expression data, there are several issues that need to be alert. The main difficulties for solving the result of optimization problem is the gene expression data is in a high dimensional with small but significant uncertainty in the original labelling's and the noise of the experimental and measurement process and the intrinsic biological variation from specimen to specimen is difficult in enhancing optimization (Ramaswamy et al., 2001). Next is gene expression data is tends to redundant, bias and confusing problem which make a classification more difficult and causing slow performance and used too much time. Lastly, for this such problems can also be posed as optimization problems of minimizing gene subset size to achieve reliable and accurate classification (Deb and Reddy, 2003). Previous research revealed that a multiclass cancer classification can be classified by SVM and among well-established and popular techniques for classification of microarray gene expression data, SVM achieve the best classification performance (George and Raj, 2011) because of the output was constructed in

hyperplane within infinite dimensional space which are linear and nonlinear.
Moreover, SVM depends on the standard of Structure Risk Minimization by taking into record of the likelihood of misclassifying yet to be seen designs for an altered however obscure likelihood conveyance of information. It utilizes a direct isolating hyperplane to make a classifier, yet it is difficult to partition a few issues in the first info space directly. Be that as it may, it can effectively change the first info space into a high dimensional component space nonlinearly, where it is minor to locate an ideal direct isolating hyperplane. The standard Support vector machine algorithm is prompts a quadratic enhancement issue with bound requirements and one direct fairness limitation. Be that as it may, when the datasets are substantial with extensive number of information focuses, the quadratic programming solvers turn out to be exceptionally troublesome, on the grounds that their time necessities.
Furthermore, memory is very reliant on the span of the preparation datasets.

Thus, this research focuses on evaluate the performance of SVM kernel function in finding the best function among linear kernel, polynomial and radial basis function (RBF) kernel.

2. Objectives

The aim of this project is to analyze the technique of Support Vector Machine (SVM) kernel function of linear kernel, polynomial and RBF kernel function that related to multiclass classification cancer. Next is to analyze the implementation of SVM classification to get the best kernel function by accuracy and computational time on multiclass cancer. Lastly, to evaluate the implementation of SVM kernel functions for multiclass cancer classification by obtaining the accuracy and time taken.

3. Methodology

In this research, the multiclass cancer classification is the main focus. Firstly, identification of problem statement and current methods of this research was identified. The objectives of the study, aim and scope were centralized according to all problems listed. Following after that is research planning such as methods and operational framework and the data set and library searching.
Secondly, on the second phase, the preprocessing data set is obtained by collecting all sets of data referencing. All the data set has gone through preprocessing step and has undergo normalization of 0 and 1. Which rescaling the data on training set for maximize the optimization for classification. The datasets are DLBCL (Shipp et al.,2002), brain tumor (Pomeroy et al., 2002) and 9_Tumors (Stuanton, 2001). Also, in the datasets obtain all the data is converted to MATLAB format (.mat). the dataset and the classes species is separated to use in classification.
Third, at the phase three, the with and without feature selection and implementation of SVM kernel function was conducted. The classifiers with different kernel functions namely as linear kernel, polynomial kernel and RBF kernel were tested for every datasets in classification after obtaining subset from feature selection and without feature selection the data immediately use in SVM classification. The rank of genes in feature selection was tested according to a certain subsets and for without selection, the actual values of samples against gene were straight away used in SVM classification while using the kernels function.
The final phase was testing and analysing. The classification was happened due to the data set is put onto SVM classifier to do their work and at the same time, the process was happened to optimize the SVM parameter and get the accuracy of the result. Here, the evaluation on data sets, analysing the data and comparing the accuracy result of classification to get the best kernel function will be conducted.

4. Result and Discussion

In Table 1, the performance analysis was executed without feature selection.

Table 1. Performance analysis of different kernel of SVM

Dataset	Kernel function accuracy (%)			Kernel function time taken to build the model (s)		
	Linear	Polynomial	RBF	Linear	Polynomial	RBF
DLBCL	**97.40**	24.68	75.32	8.349	4.062	**3.403**
Brain	**94.44**	36.67	66.67	7.596	4.184	**3.427**
9 Tumor	**85.00**	15.00	**85.00**	3.781	4.159	**3.065**

In Table 2, shows the comparison of different kernel function accuracy result, time taken to build the model as the minor analysis against dataset after ANOVA test were done and classification process was executed.

Table 2. Performance analysis of ANOVA test with different kernel of SVM classification

Dataset / No. of Genes Dataset		Kernel function accuracy (%)			Kernel function time taken to build the model (s)		
		Linear	Polynomial	RBF	Linear	Polynomial	RBF
DLBCL	10	**90.91**	79.22	80.52	6.006	7.467	**1.854**
	100	**81.82**	79.22	75.32	7.524	6.174	**1.818**
Brain	10	**77.78**	68.89	75.56	4.501	6.986	**1.591**
	100	**88.33**	78.89	66.67	8.491	4.916	**1.92**
9 Tumor	10	**88.33**	86.67	86.67	5.366	6.238	**2.426**
	100	**88.33**	76.67	85.00	5.043	3.531	**1.903**

For Table 3, the result shown is the comparison of kernel function accuracy result, time taken to build the model as the minor analysis against dataset after Signal to noise test were done and classification process was executed.

Table 3. Comparison accuries given by each algorithm

Dataset / No. of Genes Dataset		Kernel function accuracy (%)			Kernel function time taken to build the model (s)		
		Linear	Polynomial	RBF	Linear	Polynomial	RBF
DLBCL	10	90.91	75.32	**92.21**	6.038	6.789	**1.854**
	100	**96.10**	76.62	75.32	7.082	6.642	**1.818**
Brain	10	**78.89**	75.56	73.33	6.130	7.314	**1.591**
	100	**86.67**	77.78	66.67	8.898	3.423	**1.92**
9 Tumor	10	**90.00**	88.33	86.67	4.795	4.205	**1.967**
	100	**90.00**	81.67	85.00	4.607	3.084	**2.801**

Based on the result from previous section, the high accuracy obtained for the whole dataset by without feature selection is linear kernel function. Whilst, the dataset that have high accuracy of 90 percent and above is containing the informative genes which makes the accuracy higher. With this, it shows that eventhough linear kernel was set a record as best classification of two classes, it also shows that linear kernel function is good with multiclass classification. With the domination of accuracy number, shows that the function gives a good lesson to the testing dataset for multiclass problem without feature selection. Also, with the linear kernel function, it shows that these dataset are suited more with this kernel function compared to other kernel function in classifying the data. However, the time taken to build the model was evaluated but its shows that the high accuracy need more time to compute the classification. And thus, the linear kernel function was selected to be the best kernel function.

Based on the result from previous section, the high accuracy obtained for the whole dataset for both test is linear kernel. But for a Signal to noise ratio feature selection, one dataset give different number accuracy which lowering the domination of linear kernel function for the whole feature selection test classification result. The data that give one difference is DLBCL dataset with 10 samples. Whilst, the dataset that have high accuracy of 90 percent and above is actually containing the informative genes and thus the noisy data was lessen. For liner kernel function, with the domination accuracy number, shows that the function give a good lesson to the testing dataset for multiclass problem. Also, with the linear kernel function, it shows that these dataset are suit more with this kernel function compared to other kernel function in classifying the data. In addition, brain dataset have low number of informative gene when the selection of genes was made. The dataset shows clearly in the result that when the genes tested have high number, the accuracy obtained will be also high. However, the time taken to build the model was evaluated but its shows that the high accuracy need more time to compute the classification. And thus, the linear kernel function was selected to be the best kernel function.

5. Conclusion

Cancer classification is one of the challenging tasks especially using a machine learning named as Support Vector Machine (SVM). SVM is known to be good in classifying binary classes which it made difficult to deal with. Therefore, there are lots of researchers have purposed numerous attempts in classifying multiclass cancer using SVM. The main goal was to classify the data to get the best accuracy by comparing SVM kernel function by a help of with and without feature selection before the classification process. MATLAB 2013b was used in this study to point out the classification by performing accuracy of DLBCL, brain and 9 tumors dataset.

This research was conducted in to help solving the problem statement of this study as well as comparing the accuracy by with and without feature selection. In aiming to get which is the best kernel function for all the dataset tested, the experiment was performed and the results were recorded. The kernel function used for comparison is linear kernel, polynomial and radial basis function (RBF) kernel function. Also, these kernel functions give goods result in some datasets.

References

[1] Deb K., and Reddy A. R. (2003). Reliable classification of two-class cancer data using evolutionary algorithms. BioSystems, 72(1), 111-129.

[2] George G., and Raj V. C. (2011). Review on feature selection techniques and the impact of SVM for cancer classification using gene expression profile.arXiv preprint arXiv:1109.1062.

[3] Pomeroy S. L., Tamayo P., Gaasenbeek M., Sturla L. M., Angelo M., McLaughlin M. E.,and Golub T. R. (2002). Prediction of central nervous system embryonal tumour outcome based on gene expression. Nature,415(6870), 436-442.

[4] Ramaswamy S., Tamayo P., Rifkin R., Mukherjee S., Yeang C. H., Angelo M., and Golub T. R. (2001). Multiclass cancer diagnosis using tumor gene expression signatures. Proceedings of the National Academy of Sciences,98(26), 15149-15154.
[5] Shipp M. A., Ross K. N., Tamayo P., Weng A. P., Kutok J. L., Aguiar R. C., and Ray T. S. (2002). Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. Nature medicine, 8(1), 68-74.
[6] Staunton J. E., Slonim D. K., Coller H. A., Tamayo P., Angelo M. J., Park J., and Mesirov J. P. (2001). Chemosensitivity prediction by transcriptional profiling. Proceedings of the National Academy of Sciences,98(19), 10787-10792.

Protein Structure Prediction Using Robust Principal Component Analysis and Support Vector Machine

Nur Aini Zakaria a,1, Zuraini Ali Shah a,2, Shahreen Kasim b,3,*

a Department of Software Engineering, Faculty of Computing, Universiti Teknologi Malaysia, 81310 Johor Bharu, Johor, Malaysia

b Faculty of Computer Science and Technology, Universiti Tun Hussein Onn Malaysia, 86400 Parit Raja, Johor, Malaysia.

ABSTRACT

Existence of bioinformatics is to increase the further understanding of biological process. Proteins structure is one of the major challenges in structural bioinformatics. With former knowledge of the structure, the quality of secondary structure, prediction of tertiary structure, and prediction function of amino acid from its sequence increase significantly. Recently, the gap between sequence known and structure known proteins had increase dramatically. So it is compulsory to understand on proteins structure to overcome this problem so further functional analysis could be easier. The research applying RPCA algorithm to extract the essential features from the original high dimensional input vectors. Then the process followed by experimenting SVM with RBF kernel. The proposed method obtains accuracy by 84.41% for training dataset and 89.09% for testing dataset. The result then compared with the same method but PCA was applied as the feature extraction. The prediction assessment is conducted by analyzing the accuracy and number of principal component selected. It shows that combination of RPCA and SVM produce a high quality classification of protein structure

Keywords: protein structure prediction RPCA robust principal component analysis support vector machine

1. Introduction

The functional and structural annotation of protein domain is one of the important roles in bioinformatics. In this context, protein structure information plays an important information key of their structural part also the features related to the biological function (S.S. Sahu et al., 2009) such as prediction of DNA binding site, implementation of a heuristic approach to find tertiary structure, reduction of conformation search space and also characterizing the folding type of a protein or its domain. S. Zhang et al. (2012) state that the exponentially growth of newly discovered protein sequences by different scientific community caused a large gap between the number of sequenceknown and the number of structure-known proteins. Hence, there exist critical challenges to develop automated method for fast and accurate determination of the structures of proteins in order to reduce gap. Therefore, there is a compulsory to implement reliable and effective computational methods for identifying the structural class of newly discovered protein based on their primary sequences.

2. Objectives

The purposes of this research are: 1. To implement Robust Principal Component Analysis (RPCA) to determine the number of principal component. 2. To implement Support Vector Machine (SVM) for protein structure classification. 3. To evaluate the performance of RPCA and SVM based on accuracy

3. Methodology

Firstly, the current issues of protein structure prediction are investigated followed by collecting research

materials such as journals, articles, conference paper and others. The data preprocessing conducted to gain higher and better prediction success rate and system performance. It also help to minimizing error in preparation be validated by machine learning algorithm. Datasets by Ding and Dubchak (2012) filtered to remove unnecessary values and information. Research continues by applying Principal component analysis (PCA) and RPCA (Croux and Ruiz-Gazen, 2005)) algorithm to extract the essential features from the original high-dimensional input vectors. The process continued by experimenting SVM with RBF kernel using the reduced and normalized features by PCA and RPCA. The final phase is the prediction assessment of the application of RPCA and SVM by the comparison of recognition ratio compared between different methods and methods used by previous researcher. Performance testing of this research by comparing classification result of protein by overall accuracy that expressed in equation 1.

$$correctly\ recognize\ protein = \frac{correctly\ recognize\ number\ of\ query\ protein}{total\ number\ of\ protein} \quad (1)$$

4. Result and Discussion

The experiment was conducted by using three approaches in order to analyse the performance of RPCA and SVM. In order to gives a clear view on performance of RPCA, the method was compared with the PCA (the basic of RPCA) and SVM. In order to select the components that contain >60% of variance, the number of PC selected are different accordingly. Table 1 shows that number of selected PC in training dataset is lower compared to testing dataset. Table 2 shows the accuracy percentage of tested approach divided by training and testing datasets.

Table 1. Number of PC selected for classification

Feature extraction	Number of PC selected for classification	
	Training	Testing
PCA	2	3
RPCA	2	4

Table 2. Comparison of SVM, PCA + SVM and RPCA + SVM

Technique	Training Dataset Accuracy (%)	Testing Dataset Accuracy (%)
SVM	84.25	84.16
PCA+SVM	74.79	84.68
RPCA+SVM	84.41	89.09

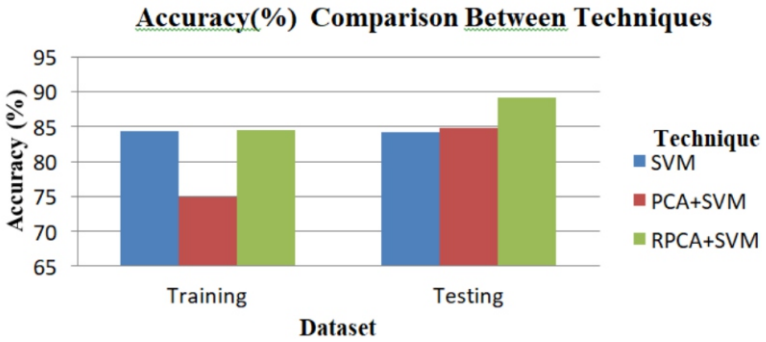


Fig 4. Accuracy comparison between techniques.

Based on this analysis, it can be assume that difference between data characteristic will influence the number of sufficient PCs required in both PCA and RPCA approach. Number of PC requires for training dataset is less then testing dataset since the size of training dataset is larger so it may contain higher information and better interpretation on features compared to testing dataset

From the results in Figure 1 it can be seen that non-extracted features technique (only SVM) gain a high percentage of accuracy (84.25% and 84.16%). However, the result can be doubt since models built on extracted features may be of higher quality, because the data is described by fewer, with more meaningful attributes. Results obtain by combination of PCA and SVM is 74.79% on training dataset and 84.68% on testing dataset. The accuracy on both datasets is quite high but stilllower than combination of RPCA and SVM technique (84.41% on training dataset and 89.09% on testing dataset). The gap seems to be higher in training dataset may because of the larger number of outliers. RPCA seems to perform the best since this method do not influenced much by outliers and its ability to detect exact fit situation.

Table 3 shows the comparison of accuracy percentage of PCA and RPCA combination with SVM. Even according by number of component, the RPCA method always seems to lead in terms of accuracy. This proves the effectiveness of RPCA approach. Table 3 also shows the increasing pattern of the accuracy for both datasets. It can be assume that higher number of PC contain much more data information lead to higher accuracy.

Table 3. Comparison of PCA+SVM and RPCA+SVM based on number of component

Number of Principal Component (PC)	Accuracy (%) for Training Dataset		Accuracy (%) for Testing Dataset	
	PCA +SVM	RPCA+SVM	PCA+SVM	RPCA+SVM
1	51.91	80.60	55.84	77.92
2	74.79	84.41	84.68	84.94
3	82.75	86.90	87.53	88.05

L. Singh, G.Chetty and D.Sharma (2012) apply the same dataset (feature vector described by Ding and Dubchak, 2001) to predict protein structure using PCA and LDA based in Extreme Learning Machine (ELM). According to the Table 4, it can be seen that proposed method used in this research shows promising results in term of accuracy obtained compare to the proposed method proposed by L. Singh, G.Chetty and D.Sharma (2012). This shows that feature extraction using RPCA and classification using SVM is an efficient method for protein structure prediction. It also shows that method proposed by L. Singh, G.Chetty and D.Sharma (2012) has drawbacks in due to the outliers and low ability in detection of exact fit situation.

Table 4. Accuracy comparison between method

Method	Accuracy (%)
LDA-ELM	77.67
PCA-ELM	82.45
RPCA-SVM	89.09

5. Conclusion

This research focus is on protein structural classification. Protein Structure classification is important for identification of protein function. As the protein structure classification n is a first and key step in protein structure prediction, it becomes an increasingly challenging task. Recently, the exponentially increase of sequence data protein cause the increasing of the requirements for reliable and effective computational method for protein structure classification. Protein structure classification is very important in bioinformatics field. Proposed feature extraction method, Robust Principal Component Analysis (RPCA) combines with Support Vector Machine (SVM) shows that data with extracted features can obtain higher accuracy (84.41% for training dataset and 89.09% for testing dataset). It also shows that RPCA works well with highly corrupted data especially dataset with outliers.

References

[1] Croux, C. and Ruiz-Gazen, A. (2005), "High breakdown estimators for principal components: the Projection-pursuit approach revisited", Journal of Multivariate Analysis, 95, 206-226

[2] Ding, Chris HQ, and Inna Dubchak. (2001), "Multi-class protein fold recognition using support vector machines and neural networks." Bioinformatics 17.4: 349-358.

[3] Singh, Lavneet, Girija Chetty, and Dharmendra Sharma.(2012) "A novel approach to protein structure prediction using PCA or LDA based extreme learning machines." Neural Information Processing. Springer Berlin Heidelberg.

[4] Li L, Cui X, Yu S, Zhang Y, Luo Z, Yang H, et al. PSSP-RFE: Accurate Prediction of Protein structure by Recursive Feature Extraction from PSI-BLAST Profile, PhysicalChemical Property and Functional Annotations." PLoS ONE 9(3): e92863, doi:10.1371/journal.pone.0092863, (2014) .

Optimization Audicor for Normal and Abnormal Heart Sounds Characteristic

Dedi Kurniadi a,1,*, Surya Yondri a, Albar a, Roza Susanti a, David Eka Putra a, Gwo-Jia Jong b

a Department of Electrical Engineering, Politeknik Negeri Padang, Padang, Indonesia
b Department Electronic Department, National Kaohsiung University of Science and Technology, Taiwan

ABSTRACT

Heart Sounds are important things in the human body that can deliver information related to the heart condition. However, a recorded signal such as PCG and ECG that getting through Audicor still contain unexpected components or noise while the recording process happens it makes the result data from Audicor cannot directly use to recognize the condition of the heart. This research presents signal processing and data analysis to suppress the noise of the heart sounds that getting while the process of recording data happens. The cleaned heart sound will be processed in feature extraction by using FFT and PCA that capable to produce the feature both of the normal and abnormal heart sounds. For the normal case, we get the data from some healthy volunteers recorded by using Audicor. While the abnormal heart sound we focus to observe the data that contain Ventricular Septal Defect (VSD) that getting from a partner hospital. As a result, feature both normal and abnormal heart sounds can be separated.

Keywords : signal processing fast fourier transform principal component analysis phonocardiograph heart sound

1. Introduction

In heart auscultation, heart sounds are the most significant cue to recognize heart condition [1]. The function of the heart is pumping the blood throughout the body and receive it again after getting the cleaning process in the lungs. More specifically the dirty blood with poor oxygen from the body that passes through the vein vessel will be transferred from the right side of the heart into the right atrium and will be transferred into the right ventricle through the tricuspid valve. Then the blood from the right ventricle will be pumped into the lung through the pulmonary valve. After that, the blood flows through the small size of the vascular vessel surrounding the air pockets in the lungs to absorb oxygen and release carbon dioxide, and then blood will flow back to the heart. On the left side of the heart, the oxygen-rich blood will transfer into the left atrium through the mitral valve, then the blood will be going to the throughout the body except lungs through aortic valve [7][8][9]. In the pumping process of the blood throughout the body, the heart will produce the sound continuously that commonly known as lub and dub. Sounds that produced by the heart is coming from the closing of the heart valves. When the tricuspid and mitral valve is closing that will produce lub sound known as first heart sounds or S1. Similarly to the dub sound or second heart sound (S2), it exists because both the aortic and pulmonic valves are closing [10].

Beside S1 and S2 also have other additional heart sounds that consist of the third heart sound or S3, fourth heart sound or S4, and murmurs. These three additional heart sounds indicate the abnormal heart sounds [11].

Heart sounds that are getting through the sounds that produce when the circulation process capable to give the information related to the condition of the patient heart [2]. In general, the heart can produce two main components in every cycle that consist of the first heart sound and second heart that is commonly

known as S3 and S4, where these extra heart sounds have a lower amplitude compared to the main component of heart sound (S1 and S2) of the heart sounds [3][4]. Beside S3 and S4 also there are systolic and diastolic murmurs. These heart sounds are recording through a portable Audicor that placing in some positions in the chase of the patient such as mitral, tricuspid,aortic, and Pulmonic. By using some positions of the chase of the patient it can get the time duration and frequency value of the heart sound in one cycle. Regarding these parameters, researchers applied their algorithms for doing some observation and decide on the condition of the heart sound signals that they are observed. Observation of the normal and abnormal heart sounds can be done through the time domain and frequency domain. P.S Vikh et al have been conducted and focused their research for detecting the normal and abnormal heart sounds by using Short Time Fourier Transform (STFT) and Continuous Wavelet Transform (CWT) [5]. A. Gharehbaghi et al created a methodology to classify heart murmurs [1]. S. Barma et al implement third heart sound (S3) detection based on non-linear signal decomposition and time-frequency localization [6].

Andrizal et al use FFT to detect the data pattern categories of the combustion engine based on exhaust emission [7]. This paper implements FFT and PCA for feature extraction on normal and abnormal heart sounds. For the normal heart sounds are getting through some volunteer that has a normal heart then those recorded signal was discussing with a physician to make sure the signal possible to use as a normal signal or vise versa. Then for the abnormal heart sounds was used HS that indicate as a Ventricular Septal Defect (VSD). VSD is one of the congenital heart disease [15] that existing a hole in the septum of the muscle wall that possible to separate between right and left chambers [16] [17] [18]. In normal condition, Rich-oxygen blood will be pumped to the aorta then transfer it to the whole body from aorta and lungs. Because of the VSD safer, the blood is not directly transferred to whole the body but some blood gets push through-hole of VSD into the right ventricle. In the feature extraction approach, firstly will be done through the preprocessing technique for the recorded heart sounds that getting from Audicor. Preprocessing technique, in this case, is using IIR high pass filterin heart sounds, and for facing the ECG signal that getting from Audicor is solved through baseline removal technique, this issue is doing by using this method because the recorded data from Audicor is not always stable for the peak value of the signal so it can difficult to implement in peak detection approach. Where this peak detection is used to obtain the signal reference that needs in the segmentation process. The aim of the segmentation process is for dividing the signal into some segments (every cycle) based on peak detection in ECG signal. After that, the feature extraction technique is implemented to get the parameter both of the normal and abnormal heart sounds.

2. Material and Method

2.1. Heart Sounds Auscultation Process

In the Auscultation process, the research purposed to implement heart sounds analysis according to heart anatomy. In normal conditions, physicians normally can measure and observe the condition of the patient hearts through a traditional way which is placing the stethoscope on the chest patient.In order to measure the sound of the patient heart, four common places normally used for auscultation process, which are mitral, tricuspid, pulmonic, and aortic as shown in the following figure [12].

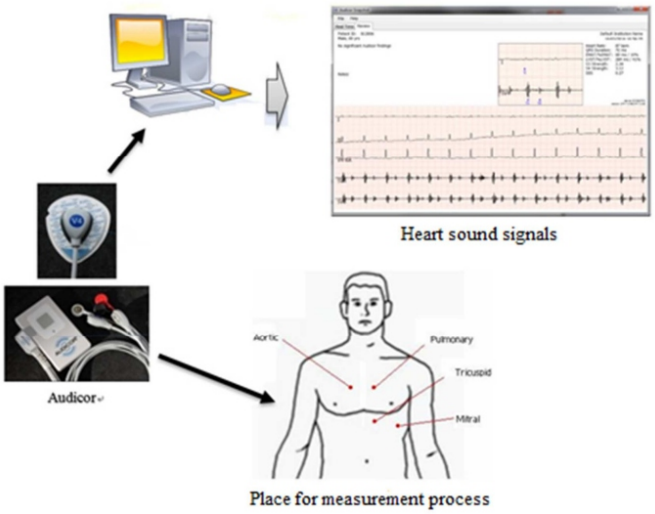


Fig 1. Measurement process of the heart sounds

Phonocardigraph is one of the most common methods implemented for heart sounds auscultation. In recording the heart sounds, this paper applied the Audicor for getting the heart sounds and Electrocardiograph (ECG) signals. The recorded signals are going to be processed and analyzed to obtain the characteristic both of normal and abnormal heart sounds. These characteristics will be discussed with an expert or physician in heart disease. Then, the recorded ECG signals are used as a parameter in the signal processing approach [13]. For supporting data, this research uses data samples both of the normal and abnormal. The data as a normal patient is obtaining through some volunteers that have a good condition in health. While the sample data that used as an abnormal patient are getting from a cooperative hospital. So, the normal and abnormal heart sound we have 20 sample data respectively.
This paper has some steps in the signal processing approach that can be used for separating the feature each of the normal and abnormal heart sounds that consist as follows:

2.2. Flowchart of The Signal Processing Approach

There are three steps in the signal processing approach that consists of preprocessing, segmentation, and feature extraction as shown in figure 3.

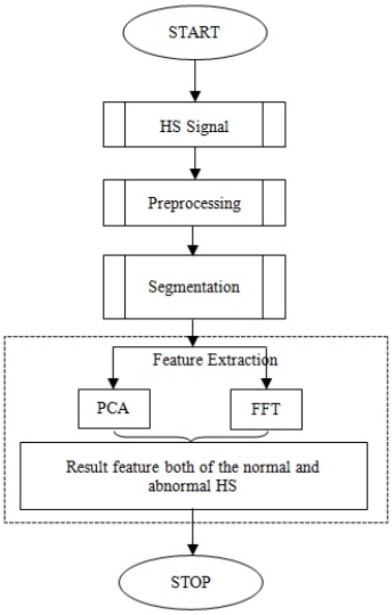


Fig 2. Flowchart of the process

The flowchart illustrates the process of the signal processing approach that is used in this research.

Firstly the data are recorded through an Audicor that will produce both of the PCG and ECG signals.

Raw data from Audicor are going to process through a preprocessing technique. In this stage, data will be cleaned from the noise to get the real data of PCG and ECG. In this case, the bandpass filter is implemented (BPF) for the PCG signal to suppress the noise that has been got when the recording data is happening. However, for ECG signal is used Baseline Removal or Baseline Wandering technique for facing the noise before implementing the peak detection.

In the stage of segmentation, this research [14] applied wavelet decomposition and reconstruction to get the detail and approximation of the frequency difference of the original PCG signal, that is capable of separating PCG signals into several parts according to peak detection and implementation of cross-correlation and normalization cross-correlation to choose the best signal that can be processed in the next step.

Feature extraction is the most important stage in this research because this stage will produce the result that makes us capable to recognize and distinguish between normal and abnormal heart sounds. This step implements the FFT and PCA.

2.3. Segmentation Process

Segmentation is the process that implements cross-correlation and normalizes cross-correlation for separating the HS signal into some parts. By using peak detection as a reference signal for the cutting process of HS signal, the research capable to choose a better signal from some segments that have been made previously. This selected signal is getting through cross-correlation as shown in equation 1 dan normalize cross-correlation in equation 2.

$$Corr\ x, y = \sum_{N=0}^{N-1} x[n]y[n] \quad (1)$$

Equation 1 illustrates the working principle of cross-correlation, where the signal x will be compared with signal y and x as a reference signal. The output number will get an indication as a similarity level, the high number considered as a high similarity. After that, we create standardize for cross-correlation known as normalized cross-correlation.

3. Results and Discussion

As mentioned earlier, this literature review will be limited to journals published in 2010 through 2018. After facing some steps in signal processing approach such preprocessing for PCG and ECG signals so it is continued with the segmentation process for heart sound signals. Where the heart sound signals will be classified into some parts based on peak detection that implements cross correlation and normalizes cross-correlation results to get the best signal that is going to be processed in the next steps.

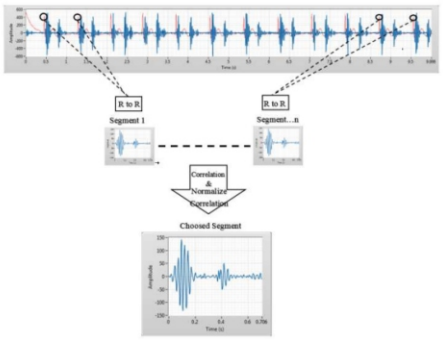


Fig 3. Segmentation Result

In feature extraction, this research use FFT to observe both of the normal and abnormal heart sounds based on the amplitude value of the heart sounds signal and this result will be used as an input in a statistical data observation on PCA. PCA is used to do the observation and analyze the different parameters both of normal and abnormal heart sounds.

Table 1. Summary Statistics (Quantitative data) for Normal HS

Variable	Obs	miss data	without miss data	Minimum	Max	Mean	Std. deviation
P1	45	0	45	0.693	14.970	5.335	3.518
P2	45	0	45	0.416	15.003	4.586	3.569
P3	45	0	45	0.814	19.422	7.102	5.316
P4	45	0	45	0.538	11.578	3.906	2.959
P5	45	0	45	0.546	6.630	2.508	1.411
P6	45	0	45	0.190	6.610	2.310	1.680
P7	45	0	45	0.156	3.310	1.438	0.826
P8	45	0	45	0.208	5.052	1.627	1.203
P9	45	0	45	0.129	6.926	1.863	1.549
P10	45	0	45	0.323	13.002	2.948	2.933
P11	45	0	45	0.095	6.057	2.527	1.773
P12	45	0	45	0.174	4.893	1.642	1.153
P13	45	0	45	0.221	8.693	3.192	2.407
P14	45	0	45	0.332	5.868	1.934	1.516
P15	45	0	45	0.289	5.086	1.547	1.022
P16	45	0	45	0.252	7.379	2.537	1.690
P17	45	0	45	0.251	6.427	1.660	1.247
P18	45	0	45	0.178	10.904	3.373	2.635
P19	45	0	45	0.377	14.714	5.392	3.839
P20	45	0	45	0.235	14.714	5.251	3.970

According to both of the normal and abnormal heart sound data that consist of 20 sample in normal condition and 20 sample data for the abnormal condition too, we can see the difference between normal and abnormal heart sound through the maximum value of summary statistic table. This condition can be more clear by using the graph to separate both of the normal and abnormal HS.

Table 1. Summary Statistics (Quantitative data) for Abnormal HS

Variable	Obs	miss data	without miss data	Minimum	Max	Mean	Std. deviation
P21	45	0	45	0.302	64.381	19.610	19.568
P22	45	0	45	0.210	15.123	4.553	3.817
P23	45	0	45	0.400	32.091	10.374	8.827
P24	45	0	45	0.324	19.769	8.127	6.246
P25	45	0	45	0.195	41.945	16.285	14.106
P26	45	0	45	0.292	45.054	12.783	14.219
P27	45	0	45	0.247	10.148	3.804	2.620
P28	45	0	45	0.357	12.813	5.136	3.613
P29	45	0	45	0.116	78.672	18.517	20.606
P30	45	0	45	0.313	16.367	6.221	4.400
P31	45	0	45	0.983	17.349	6.031	4.146
P32	45	0	45	0.625	44.381	12.304	10.921
P33	45	0	45	0.432	19.992	6.537	5.377
P34	45	0	45	0.336	38.783	11.317	11.413
P35	45	0	45	0.873	72.897	24.085	23.543
P36	45	0	45	0.119	47.955	12.990	14.219
P37	45	0	45	0.445	19.128	7.859	6.439
P38	45	0	45	0.108	40.093	11.397	12.444
P39	45	0	45	0.176	36.327	15.947	9.278
P40	45	0	45	3.267	59.105	26.926	16.118

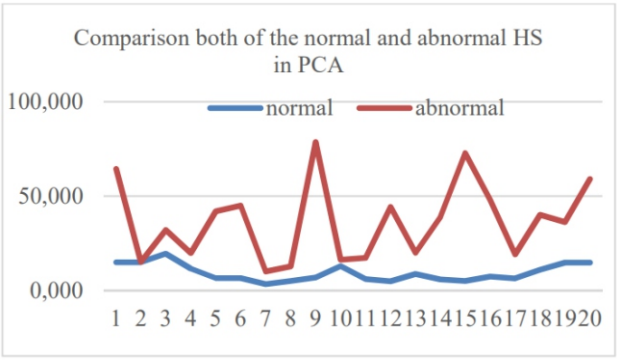


Fig 4. Graph of normal and abnormal heart sounds

Figure 5 shows the differences between normal and abnormal heart sounds. The normal heart sounds indicated by the blue color of the line and the abnormal one shows in the red line. From the tren thatshowing in this graph can give the information and separate both of the normal and abnormal heart sounds

4. Conclusion

This research implements some steps in signal processing that consists of preprocessing, segmentation, and feature extraction. In preprocessing and segmentation technique the signal will be processed to get the real signal and ready to be used as an input for the feature extraction stage. Feature extraction implements FFT and PCA to recognize characteristics both of normal and abnormal heart sounds. For the implementation process, this paper use FFT to detect the differences between normal and abnormal heart sound based on the amplitude value, and this result also will be used as an input for statistical data processing in PCA. While for the PCA itself uses maximum value on the summary statistic table of the normal and abnormal heart sounds. These two kinds of normal and abnormal data in summary statistic data value successfully show a difference parameter between normal and abnormal heart sounds.

Acknowledgment

This research is conducted through DIPA Politeknik Negeri Padang with contract number **149/PL9.1.4/PT.01.02/2019,** because of that, the author would like to say thanks to Politeknik Negeri Padang that have been funded this research.

References

[1] A. Gharehbaghi, M. Borga, B. J. Sjöberg, and P. Ask, "A novel method for discrimination between innocent and pathological heart murmurs," Med. Eng. Phys., vol. 37, no. 7, pp. 674–682, 2015.

[2] Kurniadi, D., Kung, Y.-F., Chen, Z.-H., Li, Y.-P., Hendrick, & Jong, G.-J. (2018). "Implemented the expert system of heart disease by using SVM", 2018 IEEE International Conference on Applied System Invention (ICASI).

[3] L. Hamza Cherif, S. M. Debbal, and F. Bereksi-Reguig, "Segmentation Of Heart Sounds and Heart Murmurs," Journal of Mechanics in Medicine and Biology. 2008.

[4] Kurniadi, D., Kung, Y.-F., Chen, Z.-H., Li, Y.-P., Hendrick, & Jong, G.-J. (2018). "Implemented the expert system of heart disease by using SVM". 2018 IEEE International Conference on Applied System Invention (ICASI).

[5] P. S. Vikhe, N. S. Nehe, and V. R. Thool, "Heart Sound Abnormality Detection Using Short Time Fourier Transform and Continuous Wavelet Transform," 2009 Second Int. Conf. Emerg. Trends Eng. Technol., no. 1, pp. 50–54, 2009.

[6] W. et al. Barma, Shovan; Chen, Bo-Wei; Ji, "Detection of the Third Heart Sound Based on Nonlinear Signal Decomposition and Time-Frequency Localization IF3.57 Q1," vol. 63, no. 8, pp. 1718–1727, 2016.

[7] Andrizal, R. Chadry, and A. I. Suryani, "Embedded System Using Field Programmable Gate Array (FPGA) myRIO and LabVIEW Programming to Obtain Data Patern Emission of Car Engine Combustion Categories," JOIV Int. J. Informatics Vis., vol. 2, no. 2, p. 56, 2018.

[8] "How the Heart Works - Topic Overview." [Online]. Available: https://myhealth.alberta.ca/health/pages/conditions.aspx?Hwid=tx4097abc. [Accessed: 09-Apr-2018].

[9] "(18) Video Pembelajaran : Sistem Peredaran Darah Manusia - YouTube." [Online]. Available: https://www.youtube.com/watch?v=j0t8Lif8NZc. [Accessed: 10-May-2019].

[11] "Flow through the heart (video) | Khan Academy." [Online]. Available: https://www.khanacademy.org/science/health-and-medicine/circulatory-system/circulatory-system introduction/v/lub-dub. [Accessed: 24-Apr-2018].
[12] A. Mondal, P. Bhattacharya, and G. Saha, "An automated tool for localization of heart sound components S1, S2, S3 and S4 in pulmonary sounds using Hilbert transform and Heron's formula," Springerplus, vol. 2, no. 1, pp. 1–14, 2013.
[13] Aripriharta, "The Biomedical Signal Processing Combined with Internet of Things The Biomedical Signal Processing Combined with Internet of Things," 2017.
[14] M. Nassralla, Z. El Zein, and H. Hajj, "Classification of Normal and Abnormal Heart Sounds," 2017.
[15] L. Huiying, L. Sakari, and H. Iiro, "A heart sound segmentation algorithm using wavelet decomposition and reconstruction," Proc. 19th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. 'Magnificent Milestones Emerg. Oppor. Med. Eng. (Cat. No.97CH36136), vol. 4, no. C, pp. 1630–1633, 1997.
[16] A.Wang andT. M.Bashore, Valvular Heart Disease. 2009.
[17] "Ventricular Septal Defect (VSD)." [Online]. Available: http://www.stanfordchildrens.org/ en/topic/default?id=ventricular-septal-defect-vsd-90-P01829. [Accessed: 24-Apr-2018].
[18] "Ventricular Septal Defect (VSD) in Children | Phoenix Children's Hospital Heart Center." [Online]. Available: http://heart.phoenixchildrens.org/heart-conditions/ventricular-septal-defect-vsd-children. [Accessed: 24-Apr2018].
[19] "Ventricular septal defect (video) | Khan Academy." [Online]. Available: https://www.khanacademy.org/ science/health-and-medicine/circulatory-system-diseases/acyanotic-heart-diseases/v/rn-ventricular-septal-defect. [Accessed: 24-Apr-2018]

Classification of Biomedical Literature in Hypertension and Diabetes

Nur Aniq Syafiq Rodzuana,1, Shahreen Kasima,2,*, Mohanavali Sithambranathana, Muhammad Zaki Hassana

a Faculty of Computer Science and Technology, Universiti Tun Hussein Onn Malaysia, 86400 Parit Raja, Johor, Malaysia.

ABSTRACT

Textual information gives us more clear information as it is presented using words and characters, which is easy for humans to understand. To extract this kind of information, text mining was introduced as new technology. Text mining is the process of extracting non-trivial patterns or knowledge from text documents or from textual databases. The purpose of this research paper is to perform and compare keyword extraction using statistical and linguistic extraction tools for 120 text documents related to hypertension and diabetes disease. In order to draw this comparison, RStudio, a statistical-based tool and TerMine, a linguistic-based tool have been used to demonstrate the process of extracting the specified keyword from the biomedical literature. Thus, classification evaluation using Naïve Bayes classifier is carried out in order to evaluate and compare the performance of the statistical and linguistic approaches using these tools. Experimental results show the result of the comparison and the difference between both tools in executing extraction keywords.

Keywords: classification biomedical literature hypertension diabetes

1. Introduction

Diabetes or medically termed, 'Diabetes Mellitus' is categorized as a high blood glucose level that results in the deficiency of insulin produced in the body, or the body's resistance to the effect of insulin [1]. Frequent urination, increased thirst and increased hunger are signs of high blood sugar. Diabetes mellitus has two types of categories. Type 1 diabetes is insulin-dependent diabetes (IDDM). This type occurs when there is no longer insulin or very little insulin produced by the pancreas. The other type of diabetes is non-insulin produced by pancreas or the insulin produced is not absorbed effectively by the cell in the body [4].

Over 246 million people suffer from diabetes worldwide with a majority of them being women. According to the World Health Organization (WHO) report, diabetes is ranked fifth as the fatal disease with no treatment yet to be reported and the amount of individual diagnosed from this disease is predicted to increase to over 380 million by 2025[6].

Hypertension also known as high blood pressure can be affected by many factors, such as physical inactivity, tobacco and alcohol use. There are almost one billion people who have been affected with hypertension or high blood pressure, in which two-thirds are in developing countries according to World Heart Federation. According to Centers for Disease Control and Prevention, in 2014, more than 410000 Americans had lost their lives due to the hypertension that includes 1100 deaths per day.

This is alarming because it will make the heart work harder to pump blood out to the body and it will lead to the hardening of arteries, also to stroke, kidney disease and the development of heart failure [5]. Textual information is presented by using words and characters and this will provide a lot of fine information for users. Therefore, the purpose of this research is to perform and compare keyword extraction using statistical and linguistic extraction tools from text documents related to the treatment of

of hypertension and diabetes.

2. Method

2.1. Dataset Collection

The framework of the research will be presented based on the research flow. There are six phases established before the end result can be achieved. The following Figure 1 shows the flow of this research framework.

2.2. Research Framework

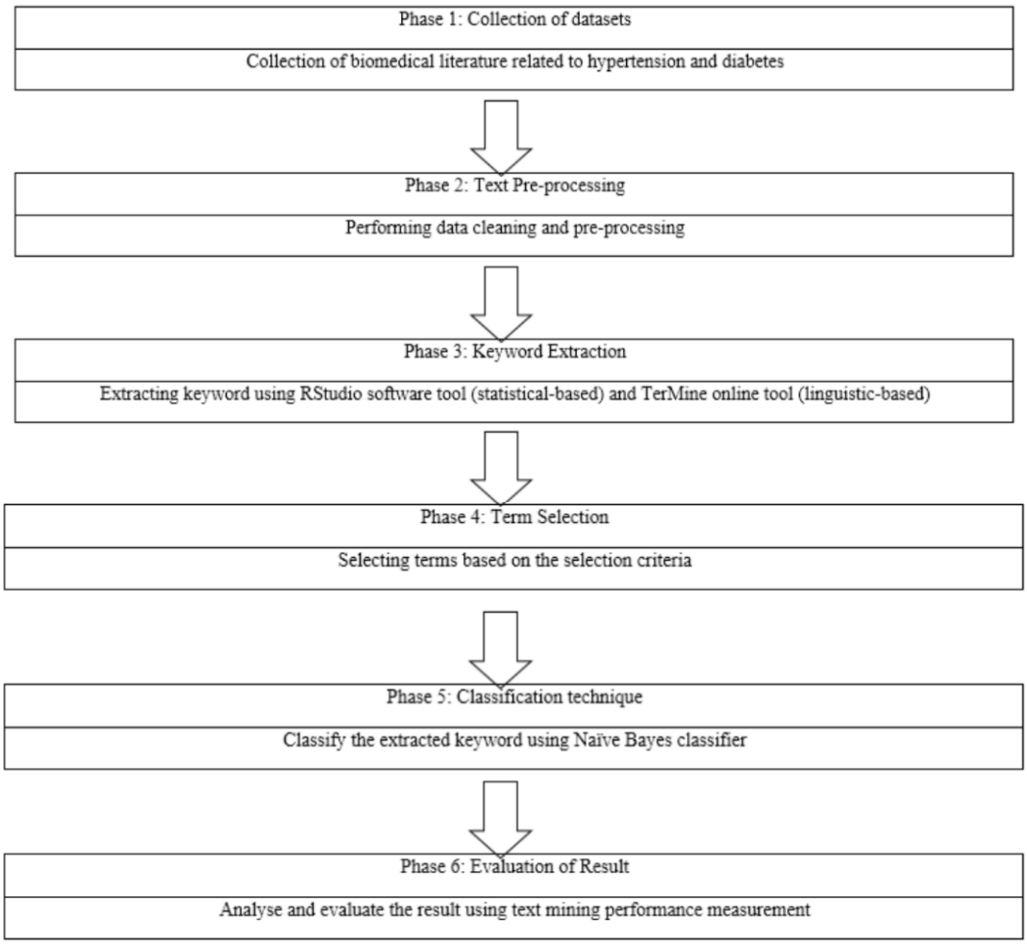


Fig 1. Research workflow

2.2.1. Phase 1: Collection of Datasets

The most important step r to initiate this research is to collect the datasets. The datasets are collected mostly from the Internet as most of the papers related to the research can be downloaded easily. The text documents used in this researchare is 60 hypertension and 60 diabetes journals and research papers collected from PubMed website, https://www.ncbi.nlm.nih.gov/pubmed/. PubMed has been chosen because PubMed comprises more than 30 million citations for biomedical literature from MEDLINE, life science journals and online books (PubMed). The documents are selected based on biomedical texts that consist keywords related to hypertension and diabetes.

The keywords used to search the biomedical literature are “hypertension” and “diabetes”. For the

keyword “hypertension”, as the first result after inserting the keyword in the search bar, 516,525 resources were shown. There are many resources but some of the documents were already outdated as they had been published since year 1990 and above. Therefore, after applying the publication dates filter for 5 years, there were 107,907 resources left. For the keyword “diabetes”, as the first result after inserting the keyword, 717,782 resources were shown. The publication dates filter for 5 years was also applied to this keyword and the results show that there were 2017,116 resources shown within the range of 5 years.

2.2.2. Phase 2: Text Pre-processing

During this phase, text pre-processing is done in order to make sure that the important keywords are included in the text. There are two types of text pre-processing involved in this research, which are statistical and linguistic pre-processing. There are many types of phases in text preprocessing, but in this research, the phases involved are data cleaning, stop word removal, stemming and part-of-speech (POS) tagging.

a. Statistical Pre-processing

The statistical pre-processing was carried out using the RStudio tool. The tool can be downloaded from the website https://www.rstudio.com/products/rstudio/download. After installing the RStudio, the R tool must be installed in the computer in order to make sure that RStudio is functioning.; the tool can be downloaded from the website https://cran.rproject.org/bin/windows/base/.

In the statistical pre-processing, the phases involved are data cleaning, stop word removal and stemming. The data cleaning phase includes a few steps such as characters conversion into lowercase, numeric removal, punctuation removal and whitespace removal. Stop word removal is a phase where the number of common words used in the text document is reduced. Next, for stemming, words in the text documents will be classified in terms of their root or stem words. All the three phases were done by running the command in Rstudio.

b. Linguistic Preprocessing

This process helped in finding named entities and improved the selection of nouns or other important words from a document [7]. All related terms were identified by applying POS tagging, extracting word sequences of adjectives or nouns and stop-list. The linguistic pre-processing was done using the TerMine tool.

2.2.3. Phase 3: Keyword Extraction

After the pre-processing for the biomedical literature was complete in the previous phase, this phase is all about the keyword extraction on the text documents. The tools in this phase are the same tools used in the previous phase which are RStudio and TerMine.

a. Statistical Approach using Rstudio

RStudio is an integrated development environment (IDE) for R. RStudio includes a console, a syntax-highlighting editor that supports direct code execution, and tools for plotting, history, debugging and workspace management. RStudio provides the most widely used open source and it is enterprise-ready and it can be run on the desktop or in a browser that was included with RStudio Server or RStudio Server Pro. Term frequency is used as the method of statistical feature in order to measure the keyword. Term frequency indicates how often a word occurs in a text document.

b. Linguistic Approach using TerMine

TerMine is an online text mining tool by The National Centre for Text Mining (NaCTeM). The TerMine demonstrator combines the c-value multiword term extraction and AcroMine acronym recognition. The demonstration system will explain the input resources by recognising the multiword term by c-value and acronym recognised by AcroMine.

The method of the linguistic feature used in this research to measure the keyword is c-value. By using c-value, the result from the keyword extraction will be represented with the c-value score. C-value combines the linguistic and statistical analyses as it is a domain-independent method for automatic term recognition (ATR).

2.2.4. Phase 4: Term Selection

During this phase, the selection is completed by selecting through exact selection criteria. For the statistical approach, the term frequency higher than, or equal to 30, is selected, and for the linguistic approach, the c-value score higher than, or equal to, 3 is selected. Term selection is used because the range of the targeted extracted keyword is in the range of 80 to 120. Term selection can help in improving the prediction performance in classifying the data as it removes unnecessary features within the document [3].

2.2.5. Phase 5: Classification Technique

Weka tool accepts only the Arff file format. Hence, ArffViewer in Weka is available to convert other file formats into the Arff file format. Classification is completed by choosing the right classifier in Weka. In this research, the classifier used is Naïve Bayes classifier. The method used to train the data is the K-fold cross-validation.

Naïve Bayes is a simple Bayesian Network that assumes on finding nodes that are restrictively independent of each other [10]. The term represents that there is restrictive independence among the features or attributes. The probability parameters are predicted from the training data where the parameter predicted from the data is completed by using maximum likelihood estimation [9].

K-fold cross-validation is a good evaluator for acquiring the error rate of a learning algorithm [8]. This method is one of the most popular and practical methods because it is simple and it has an obvious universality. It is also used to evaluate the probability of an evaluator [2].

2.2.6. Phase 6: Evaluation of Result

The final phase of this research is to evaluate the result. After producing the result, the result will be analysed and evaluated using performance measurement. The measurement of the performance of the classification task in Weka will be using the precision, recall and F-measure. Precision measures the exactness of a classifier. A higher precision means less false positives, while a lower precision means more false positives. Recall measures the completeness, or sensitivity of a classifier. Higher recall means less false negatives, while lower recall means more false negatives. Precision and recall can be combined to produce a single metric known as F measure, which is the weighted harmonic mean of precision and recall. The main advantage of using F-measure is that it is able to rate a system with one unique rating.

3. Result and Discussion

True Positive (TP) gives information on how often the data that correctly classified the document is related to the diseases. False Positive (FP) refers to how often the data that classified the document was related to the diseases when there was no relation at all. Next, True Negative (TN) is to see how often the

data that correctly classified the document was unrelated to the diseases. Finally, False Negative (FN) deals with how often the data that classified the document was not seemingly related to the diseases but in fact, was related.

Figure 1 shows the result of performance measurement for both statistical and linguistic based tools. For the statistical-based tool, the average precision value is 0.588, the recall value is 0.592and the F-measure value is 0.590. There is less difference between precision and recall. This is probably because the values of false positive (FP) and false negative (FN) are close to each other.

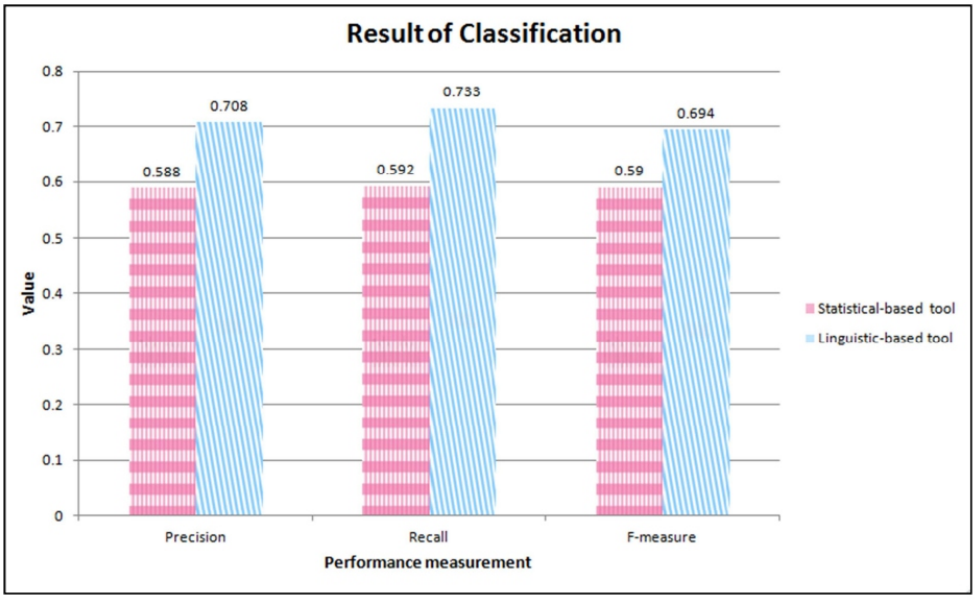


Fig 2. Performance measurement result

Table 1. Result of database hypertension and diabetes using Naïve Bayes

Approaches	Accuracy (%)	Precision (%)	Recall (%)	F-measure (%)
Statistical (RStudio)	60.3	58.8	59.2	59.0
Linguistic (TerMine)	72.4	70.8	73.3	69.4

4. Conclusion

For the linguistic-based tool, the average precision value is 0.708, the recall value is 0.733, and F measure is 0.694. The result of the linguistic-based tool is obviously higher compared to the result of the statistical-base tool. The higher the F-measure value, the better the predictive classification procedure. A score of 1 means the classification procedure is perfect while the lowest possible F measure is 0. In this research, the value of 0.694 is near-perfect because the F-measure value is nearest to 1 instead of the value of 0.90. Therefore, the result from the linguistic-based tool is better than the statistical-based tool.

References

[1] Ali, R., Hussain, J., Siddiqi, M. H., Hussain, M., and Lee, S. (2015). H2RM: A Hybrid Rough Set Reasoning Model for Prediction and Management of Diabetes Mellitus, 15921–15951.

[2] Arlot, S., and Celisse, A. (2009). A survey of cross-validation procedures for model selection, 4, 40–79. doi: 10.1214/09-SS054.

[3] Chandrashekar, G., and Sahin, F. (2014). A survey on feature selection methods. Computers and Electrical Engineering, 40(1), 16–28. doi: 10.1016/j.compeleceng.2013.11.024.

[4] Gülçin Yıldırım, E., Karahoca, A., and Uçar, T. (2011). Dosage planning for diabetes patients using data mining methods. Procedia Computer Science, 3, 1374–1380.
[5] Holland K., (2017), Everything You Need to Know About High Blood Pressure (Hypertension). Retrieved from http://www.healthline.com/health/high-blood-pressure-hypertension.
[6] Iyer, A., S, J., and Sumbaly, R. (2015). Diagnosis of Diabetes Using Classification Mining Techniques. International Journal of Data Mining and Knowledge Management Process, 5(1), 01–14.
[7] Jurafsky, D., and Martin, J. H. (2016). Part-of-Speech Tagging. In Speech and Language Processing. Retrieved from http://en.wikipedia.org/w/index.php?title=Part-of-speech_tagging&oldid=550410494.
[8] Kale, S., Kumar, R., and Vassilvitskii, S. (2011). Cross-Validation and Mean-Square Stability.
[9] Pineda A. L., Yea Y., Visweswarana S., Coopera G. F., Wagnera M. M., and Tsuia F., J Biomed Inform. (2015) December ; 58: 60–69. doi:10.1016/j.jbi.2015.08.019.
[10] Spasić, I., Greenwood, M., Preece, A., Francis, N., and Elwyn, G. (2013). FlexiTerm: a flexible term recognition method. Journal of Biomedical Semantics, 4(1), 27.

Instructions for Authors

Essentials for Publishing in this Journal

1 Submitted articles should not have been previously published or be currently under consideration for publication elsewhere.

2 Conference papers may only be submitted if the paper has been completely re-written (taken to mean more than 50%) and the author has cleared any necessary permission with the copyright owner if it has been previously copyrighted.

3 All our articles are refereed through a double-blind process.

4 All authors must declare they have read and agreed to the content of the submitted article and must sign a declaration correspond to the originality of the article.

Submission Process

All articles for this journal must be submitted using our online submissions system. http://enrichedpub.com/ . Please use the Submit Your Article link in the Author Service area.

Manuscript Guidelines

The instructions to authors about the article preparation for publication in the Manuscripts are submitted online, through the e-Ur (Electronic editing) system, developed by **Enriched Publications Pvt. Ltd**. The article should contain the abstract with keywords, introduction, body, conclusion, references and the summary in English language (without heading and subheading enumeration). The article length should not exceed 16 pages of A4 paper format.

Title

The title should be informative. It is in both Journal's and author's best interest to use terms suitable. For indexing and word search. If there are no such terms in the title, the author is strongly advised to add a subtitle. The title should be given in English as well. The titles precede the abstract and the summary in an appropriate language.

Letterhead Title

The letterhead title is given at a top of each page for easier identification of article copies in an Electronic form in particular. It contains the author's surname and first name initial .article title, journal title and collation (year, volume, and issue, first and last page). The journal and article titles can be given in a shortened form.

Author's Name

Full name(s) of author(s) should be used. It is advisable to give the middle initial. Names are given in their original form.

Contact Details

The postal address or the e-mail address of the author (usually of the first one if there are more Authors) is given in the footnote at the bottom of the first page.

Type of Articles

Classification of articles is a duty of the editorial staff and is of special importance. Referees and the members of the editorial staff, or section editors, can propose a category, but the editor-in-chief has the sole responsibility for their classification. Journal articles are classified as follows:

Scientific articles:

1. Original scientific paper (giving the previously unpublished results of the author's own research based on management methods).

2. Survey paper (giving an original, detailed and critical view of a research problem or an area to which the author has made a contribution visible through his self-citation);

3. Short or preliminary communication (original management paper of full format but of a smaller extent or of a preliminary character);

4. Scientific critique or forum (discussion on a particular scientific topic, based exclusively on management argumentation) and commentaries. Exceptionally, in particular areas, a scientific paper in the Journal can be in a form of a monograph or a critical edition of scientific data (historical, archival, lexicographic, bibliographic, data survey, etc.) which were unknown or hardly accessible for scientific research.

Professional articles:

1. Professional paper (contribution offering experience useful for improvement of professional practice but not necessarily based on scientific methods);
2. Informative contribution (editorial, commentary, etc.);
3. Review (of a book, software, case study, scientific event, etc.)

Language

The article should be in English. The grammar and style of the article should be of good quality. The systematized text should be without abbreviations (except standard ones). All measurements must be in SI units. The sequence of formulae is denoted in Arabic numerals in parentheses on the right-hand side.

Abstract and Summary

An abstract is a concise informative presentation of the article content for fast and accurate Evaluation of its relevance. It is both in the Editorial Office's and the author's best interest for an abstract to contain terms often used for indexing and article search. The abstract describes the purpose of the study and the methods, outlines the findings and state the conclusions. A 100- to 250-Word abstract should be placed between the title and the keywords with the body text to follow. Besides an abstract are advised to have a summary in English, at the end of the article, after the Reference list. The summary should be structured and long up to 1/10 of the article length (it is more extensive than the abstract).

Keywords

Keywords are terms or phrases showing adequately the article content for indexing and search purposes. They should be allocated heaving in mind widely accepted international sources (index, dictionary or thesaurus), such as the Web of Science keyword list for science in general. The higher their usage frequency is the better. Up to 10 keywords immediately follow the abstract and the summary, in respective languages.

Acknowledgements

The name and the number of the project or programmed within which the article was realized is given in a separate note at the bottom of the first page together with the name of the institution which financially supported the project or programmed.

Tables and Illustrations

All the captions should be in the original language as well as in English, together with the texts in illustrations if possible. Tables are typed in the same style as the text and are denoted by numerals at the top. Photographs and drawings, placed appropriately in the text, should be clear, precise and suitable for reproduction. Drawings should be created in Word or Corel.

Citation in the Text

Citation in the text must be uniform. When citing references in the text, use the reference number set in square brackets from the Reference list at the end of the article.

Footnotes

Footnotes are given at the bottom of the page with the text they refer to. They can contain less relevant details, additional explanations or used sources (e.g. scientific material, manuals). They cannot replace the cited literature.
The article should be accompanied with a cover letter with the information about the author(s): surname, middle initial, first name, and citizen personal number, rank, title, e-mail address, and affiliation address, home address including municipality, phone number in the office and at home (or a mobile phone number). The cover letter should state the type of the article and tell which illustrations are original and which are not.