A Hybrid Neural Network Approach to Classify Gene Datasets

Immaculate Mercy A

PG & Research Department of Computer Science, A.V.V.M Sri Pushpam College Poondi, (Autonomous),Thanjavur

Affiliated to Bharathidasan University, Tiruchirapalli, Tamil Nadu, India

Chidambaram M

PG & Research Department of Computer Science, Rajah Serfoji Govt. Arts College (Auto), Thanjavur

Affiliated to Bharathidasan University, Tiruchirapalli, Tamil Nadu, India

ABSTRACT

The Classification and prediction of gene expression datasets can be achieved in a more appropriate manner by using the Neural Network coupled with the Modified Naïve Bayes classification algorithms. The ENNMNBC (Enhanced Neural Network coupled Modified Naïve Bayes Classification) brings out a classification and prediction of sequence Datasets in the domain of Biological and Medical Engineering. The proposed hybrid approach serves as a harness for the Modified Naïve Bayes Classifier towards handling better decisions by training the datasets using the Feed forward Neural Network. Also this approach provides a generalization for the candidate sequences .The proposed hybrid approach tries to overcome the over-fitting and the under-fitting problems which aids the domains of interest for working with the learned classifiers and predicting them. The performance measures obtained using this approach provides a solution for better predictive models. Comparing the computational costs with the existing methodologies it has considerable minimized thereby providing a better approach over the existing algorithms. The choice of implementing the threshold values produces the better probabilistic values. Moreover the features are better selected and pruned using the Sequence Generation algorithm.

Keywords: Neural Network, Naïve Bayes Classifier, Sequence Generation, Gene Sequences, Candidate Sequences.

1. INTRODUCTION

The Biological and the Biomedical Engineering domains have been formulating various methods towards drug discovery, analysis of Gene, Prognosis and diagnosis of diseases either by using the Protein-Protein interaction datasets or the publicly available gene datasets which may be curated or non-curated. Although various studies and researches has been happening in these domains still there has been found some gaps in the studies as they have trained only a small range of datasets. The traditional algorithms were all proved to be efficient when using the KNN, SVM or the Naïve Bayes method. All these methods proved to be on the front in terms of accuracy as they were sufficed to only small or a finite three or four Gene Classes of datasets. It proved to be a failure when the datasets where scaled as these traditional methods were not trained for working on larger scale. The very nature

of working with gene expression datasets is that they are high dimensional in nature. This high dimensionality has posed a threat on all the existing methods. Also the computational costs seemed to be on the rise. In the scope of eliminating the shortcomings of the existing approaches the hybrid method has been implemented.

The earlier methods suggested a two level approach of working with these datasets. The hybrid method that has been implemented is best suited for the imbalanced datasets likely the supervised, semi-supervised or the non- supervised datasets. As the Gene datasets are acquired from various publicly available repositories they are of different natures and follow a different structural pattern. The traditional approach goes through a Data Pre-processing Phase and the Classification phase. The data preprocessing mainly comprises of the feature selection and the sampling approach.

This work presents a hybridization of feed forward neural network and the Modified Naïve Bayes classification mechanism. The Modified Naïve Bayes classification proceeds by way of finding the minimum standard deviation for each range of instance.

2. EXPERMENTAL

The framework for the work proposed is as follows :

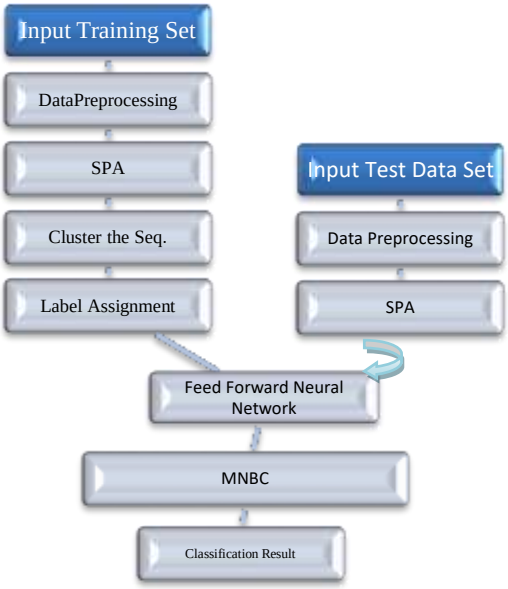

Figure 1 Work Flow

As the information taken for the examination is the DNA Splice intersections, vital planning and cleaning of information should be finished. Graft intersections are areas on strings of DNA or RNA where pointless segments are evacuated when proteins are made. After the graft, an area, known as the intron, is evacuated and the rest of the segments, known as the exons, are consolidated. The dataset

comprises of groupings of DNA that contain either the piece of the DNA held subsequent to joining, the part that was spiced out or not one or the other. Distinguishing the groupings of DNA assumes a fundamental job as the previous AI calculations represented a danger on time of calculation. To work with the difficulties that have been presented in the previous techniques another work process design has been conceived which not just aides in better cleaning and planning of information yet helps in a superior characterization strategy.

The pre-preparing is finished by the methods for Candidate Sequence Generation. The information is the DNA successions and the yield is the Candidate Set. The methodology for the Candidate Generation is as per the following:

The quality groupings are taken care of into the framework. The size of the Gene groupings is taken to be as N. The separation is figured on the quality successions which extricate a Distinct Component from the datasets.

For all the datasets in the examination the dataset is broken into singular Gene parts. On the off chance that the separation figured doesn't contain the current segment it is included into the segment list. After the particular segments are separated another competitor list should be created. This is accomplished by figuring the help and the certainty measures. Edge esteem is set for the competitors. In light of the set limit esteem the competitor grouping is produced. The given dataset is broken into unmistakable quality parts in order to help during the time spent distinguishing if any characters other than the required EI, IE stops are available. The separating of these successions helps in recognizing unmistakable segments, and if the arrangement doesn't contain the necessary parts it is added to register separation set. This gives a completely removed particular segment set. This returns by the method of computing the help and certainty measures for the applicant list. So as to achieve precision over the help and certainty edge esteem is set. In view of the edge esteem the up-and-comer grouping is produced. The acquired rundown is along these lines refreshed.

The Sequential Pattern Algorithm (SPA)

For each dataset in the input sequence

Set *Seq* to be the Gene Sequence of the user (U_j)

Break Seq by Cd_{thresh}

$$Ct=\begin{cases}\text{Increment } Ct \text{ if } Seq_t \text{ found in } cd_{list} \\ Ct \qquad\qquad Otherwise\end{cases}$$

$$W_t=ct/\text{Size } (cd_{list})$$

Let Seq_f be the individual Sequence

Let P_{seq}(f) be the $Cand_{seq}$ in Seq_F

Calculate occurrence $O_r(P_i)$ in cd_{list}

Support (Sp_f) =$O_r(P_i) + W_t+W_{ct}$

Compute the confidence for each Candidate

Set threshold (Th_x) for Pruning the sequences

Compute min confidence max confidence of the sequence

Mi_{cf}**And**Mx_{cf}

Th_x**=** **(**Th_x**%*** Mi_{cf}**)+**Mi_{cf}

If $Conf_x$**<**Th_x Pr_x**←** **Add**seq_x **to Pruned list**

The SPA Algorithm assumes a significant job which clears the exact way for order process. The calculation accepts the contribution as Gene Sequences and competitor set and the normal yield is the consecutive example. This calculation takes a shot at both the test and the prepared information. The initial step is the weight estimation where the weight is evaluated for all the up-and-comers and all the records. The grouping is broken by up-and-comer edge. This is processed by the condition

$$Ct=\begin{cases}\text{Increment } Ct & \text{if } Seq_t \text{ found in } cd_{list}\\ Ct & Otherwise\end{cases}$$

If the examined sequence is already found in the sequence set, the Ct is incremented by 1, else the current Ct is taken.

The weight is computed using the equation

$$W_t=ct/\text{Size}\ (cd_{list})$$

The obtained Ct is divided by the size of the candidate list that was generated in the previous phase. As there are three classes namely EI, IE and N the class weights needs to generated for each sequence which is given by W_{ct}. X is taken as the size for each Gene sequence. The no. of sequences available needs to be computed which is given by Y. The individual sequence is given by Seq_f. $P_{seq}(f)$ is taken as the candidate sequence in Seq_f. The occurrence of the sequence in the candidate list is found out using the support factor. The support factor is calculated using the formula

$$Sp_f = O_r(P_i) + W_t + W_{ct}$$

Where W_t is the weight of the candidates and W_{ct} is the class weight. Based on the support value achieved the next task is to compute the confidence and Lift. The confidence is computed using the equation

$$Conf_x=\sum_{i=1}^{m}(Sp_x + Sp_y)/Sp_x$$

Where Sp_x is the support count of individual gene sequence and Sp_y is the support count of the taken gene sequence. To have more accuracy over the extracted data the lift measure is calculated using the equation

$$lift_x = \sum_{i=1}^{m}(Sp_x + Sp_y)/(Sp_x * Sp_y)$$

After finding the confidence and support values for the sequences the data needs to be pruned as it has some irrelevant data. A threshold value is set for pruning the sequence. Based on the threshold the minimum confidence and maximum confidence of the sequence is computed. The threshold is set using the formula

$$Th_x = (Th_x\%* Mi_{cf})+Mi_{cf}$$

If the confidence value for the given sequence is lesser than the threshold value the sequence is added into the pruned list.

Clustering Algorithm

Let C be the Number of clusters

Let Ch_H be the initial cluster heads which are set to the clusters

$count_f$=$T_{count}(Seq_f)$

Add Mx_{ct} to Ch_H // cluster head list

Compute $Av_{ct} = Mx_{ct} - Mx_{ct}* (Av_{th})$

Add Av_{ct} to Ch_H // cluster head list

Compute$Mi_{ct} = Mx_{ct} - Mx_{ct} * (Av_{th})$

Add Mi_{ct} to ch_H

For all the sequence compute the distance as

$D_{I,J}$←Distance ($Conf_i$,$Conf_j$)

I←index (min($D_{I,J}$))

Update Seq_i to ch_{Hj}

The input for the Gene Clustering algorithm is the sequential patterns which have been extracted as an output of the SPA algorithm. The desired output is the Clustered sequence. *C* is taken to be the number of Clusters and Ch_H is the initial cluster head which are set to the clusters. The cluster head selection is done by counting the sequences. The maximum count and the minimum count of the sequences are taken as Mi_{Ct} and Mx_{ct} respectively. The maximum count Mx_{ct} is added to the cluster head list Ch_H. Based on the maximum count the average count is calculated. In the same manner minimum count is computed with respect to the maximum and the minimum count. This minimum count is added to the cluster head list. The frequent and the infrequent sequences are updated to the cluster head list. For this computation the distance is calculated based on the confidence of I and j values. The minimum index value is updated to the sequence list, which is then updated to the cluster head list for each and every candidate list j.

These clustered sequences are then fed into the Multi layer feed forward neural network. Using a multilayer feed forward neural network multiclass classification is relatively straightforward.

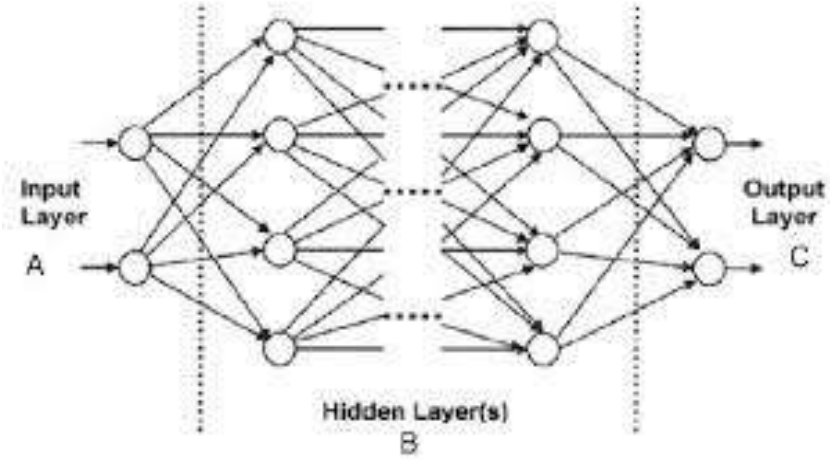

Figure 2 Architecture of Multilayer Feed Forward Neural Network

A network for binary classification only has a single output that estimates the probability that an input pattern belongs to the "yes" class, i.e., $t_i = 1$. One approach is to use multiple network outputs, one for each of the *C* classes. Using this approach, the *j*th output for the *i*th training pattern, t_{ij}, is the estimated probability that the *i*th pattern belongs to the jth class, denoted by $\hat{t}_{ij}$. An easy way to estimate these probabilities is by the useof logistic activation for each output. This ensures that each output satisfies the univariate probability requirements, i.e., $0 \le \hat{t}_{ij} \le 1$.
However, since the classification categories are mutually exclusive, each pattern can only be assigned to one of the *C* classes, which means that the sum of these individual probabilities should always equal

However, if each output is the estimated probability for that class, it is very unlikely that $\sum_{j=1}^{C} \hat{t}_{ij} = 1$. In fact, the sum of the individual probability estimates can easily exceed 1 if logistic activation is applied to every output.

Modified Naïve Bayes Classifier

The output from the Neural network now is the domain knowledge of the Modified Naïve Bayes Classifier. The modified Naïve Bayes classification aims at classifying the instance, where the clustered sequences are given as input. The size of the total No. of sequences is taken as S. The training set and the Test set sizes are initialized which is given by $(tr_s) and\ (ts_s)$ respectively. The no. of classes to be identified is given by N. For each of the classes the count of the feature set is updated into cnt_i, where *i* specify which class it belongs to. The total count is then computed. The probabilistic component computation is performed for each class. The main feature of MNBC is that it proceeds by the way of finding the minimum standard deviation for each range of instance. Within each data of the testing set and within each class, the algorithm is said to get the set of attributes for each sequence.

The number of data in a given range is assigned to Px_i if the $F_i - sd \geq data$ and $F_i + sd \leq data$. The total no. of features obtained is given by Pt_i.

The probability for each feature extracted within each class is given by

$$Px(F_i)=\prod_{i=1}^{size(F_i)} \frac{Px_i}{Pt_i}$$

The probability that a given sequence falls within a class is given by

$$Px(cl_i)= Px(F_i) * P_{comp}(i)$$

Let the three classes be given as cl_1, cl_2 and cl_3 respectively.

If the probability of Class 1 is greater than the probability class 2 and class 3 the class label is set to cl1. Otherwise if the probability of class2 is greater than the probability of class1 and class3, then the class label is set to 2, else it is set to 3.

Even though the consequences are highly correlated the MNBC offers a greater improvement over the Naïve Bayes Classification and the Naïve Hubness Classifier methods. This is in accordance with the results that have been achieved, though the amount of data taken for the analysis is highly scalable.

3. RESULTS AND DISCUSSION

The dataset that has been taken for the work is the Gene arrangement dataset for the Species Homo Sapiens. The information in its crude structure can't be utilized as it needs to grafted and the vital cleaning and planning of information should be considered for the work. After the planning we accomplish at a dataset which is joined. The Gene articulation datasets differs from stage to organize in a human body. Graft intersections are areas on strings of DNA or RNA where pointless segments are expelled when proteins are made. After the graft, a segment, known as the intron, is evacuated and the rest of the segments, known as the exons, are combined. Breaking down these segments of information is exceptionally tedious. There are many Machine learning calculations yet the technique proposed in this paper demonstrates to have a higher productivity and precision over the current grouping calculations. The dataset considered for this work comprises of successions of DNA that contain either the piece of the DNA held in the wake of grafting, the part that was joined out or not one or the other. The fundamental center is to recognize and arrange them to which case it has a place with.

There are three classes of information in particular

1. EI
2. IE
3. N

EI is the Exon – Intron Boundary, IE is the Intron – Exon Boundary and N has a place with neither of the classes. EI are called as givers and IE are called as acceptors. An EI limit is characterized by a short grouping around the limit and the nonappearance of a "stop" codon on the Exon side of the limit. The Introns are un-deciphered mediating arrangements in mRNA. The natural affirmation of expectation is

quite often vital. With such a large number of genomes being sequenced , it is imperative to have the option to recognize qualities and the signs inside and around qualities computationally.

A Sample of the datasets that was taken for the examination is given beneath. The Analysis has been made for three distinctive example sizes and the examination was done for different estimates like F1-Score for Donor, Prediction Accuracy, and Accuracy for hc19, Accuracy for Hg-38 and Accuracy factor over various example sizes.

After the planning of the information the applicant backing and certainty are registered for each unmistakable sets. This is then trailed by the method of figuring the certainty, backing and lift for every one of the classes. The particular sets are gotten by choosing an edge an incentive in the scope of 4 to 6. The edge has been set to these qualities as the arrangements that have been shown up have been taken consideration to see that their complete length is 60. Here we could state that the Gene hushing can be accomplished somewhat.

The applicant set and the Gene successions are taken care of into the framework as contribution for the EGSP procedure. For every up-and-comer grouping in the given arrangement the event of the succession is determined. The Support check is determined with the event of the grouping and the weight and competitor weight of the information. To additionally upgrade and extemporize the procedure the certainty and lift are estimated for each grouping with its help esteem. When the certainty and lift measures have been registered, the groupings should be pruned as it might contain IE stops. We are worried in just the EI codons and not IE codons. The yield gives us the Extracted Sequential example from the incessant groupings.

The removed examples at that point should be grouped. The normal, least and most extreme help check is determined for each groups and added to the bunch head. To see if the arrangements fall into the predetermined group head, the separation is determined dependent on the certainty estimates which were shown up before. The base separation reveals to us that the given succession can be categorized as one of the group heads. These bunched groupings should be characterized utilizing the Modified Naïve Bayesian arrangement calculation. The sources of info taken are the grouped groupings, testing information and the preparation information. The highlights should be removed from these groupings and a rundown of list of capabilities is made. The informational index comprise of three classes , where a portion of the records are unlabeled , these unlabelled classes are classified as N as talked about before and these datasets needs to fixed either into the EI or the IE classes. The probabilistic segments for every one of these classes are processed. To limit the mistake esteems and distinguish them precisely an insignificant standard deviation is registered for each class. The likelihood part for each component ios processed utilizing the recipe

$$〖Px(F〗 _i)=∏_(i=1)^size(F_i)▒ 〖Px〗 _i/〖Pt〗 _i$$

The typical Bayesian grouping follows by the method of summarizing the probabilistic qualities, yet in this proposed work a result of these probabilities is figured as the information acquired from the past calculation was of the bunched position. In view of the figured likelihood the segments should be fixed

into the separate classes. The correlation of qualities is finished with all the three classes and the new names are set. In this manner the grouping has been shown up effectively utilizing the EGSP and MNBC calculations.

The effectiveness and the exactness measures are upheld by the method of performing different factual tests on the current strategies and the proposed techniques. The F1 score is utilized to give a weighted normal on the exactness and review thinking about the bogus positives and the bogus negatives. Despite the fact that a lopsided class dissemination was watched the F1 score has given a better over the current techniques. The table underneath shows the correlation of F1 score on different existing strategies and the proposed techniques.

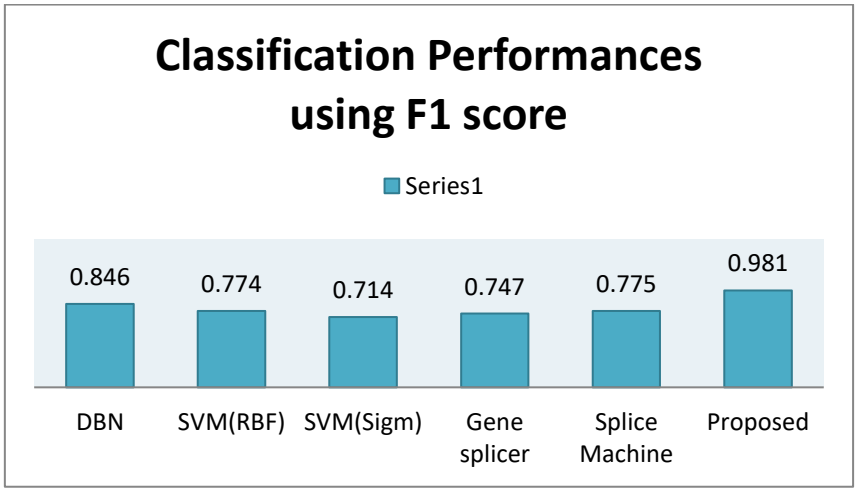

Figure 3 Classification Comparison

The order exhibitions have been looked at utilizing different classifiers and is contrasted and the proposed techniques. As can be seen unmistakably from Table 1 of qualities and the diagram it is seen that the proposed philosophy has picked up in its F1 score esteem well over the current techniques like DBN, SVM(RBF), SVM(Sigma), Gene Splicer, Splice Machine. This gives us the privacy measure on the proposed technique and sure is a banality to the arrangement strategy.

Applying this strategy further improves that it is sure to contact the characteristic of 1 on the F1 score esteems. Contemplating of how DBN functions with, the datasets could be prepared covetously each layer in turn. In spite of the fact that it demonstrates to give a superior order it has been seen that it functions admirably with unaided information. In any case, the dataset that has been taken for the investigation includes directed, semi-regulated and un-administered. Relatively from the Table 1 we could see that the worth acquired from DBN is 0.846 and that from the MNBC is 0.981. The various classifiers like the SVM (RBF), SVM (sigma), Gene Splicer and Splice Machine it has been seen from the chart that definitely there is a dunk in the presentation contrasting with the proposed classifier. This is accomplished by registering the base standard deviation for the each group head and this separation measure helps in processing the probabilistic parts for each class.

The yield achieved gives us a view over the better classifier. This shows an unmistakable division that the proposed classifier stands apart from the various classifiers.

4. CONCLUSION

The Hybrid methodology of the Feed forward Neural Network and MNBC calculations were proposed to have an edge over the current order calculations, where the current calculations demonstrated to function admirably for littler examples. In any case, with regards to the point of versatility they have anyway indicated a reasonable plummet in the qualities accomplished. The edge esteems that have been set for the competitor age offered an approach to achieve precision for the further handling. The pruning of the successions have added to the reality for decidability factor by methods for registering the base and the most extreme certainty gauges .The bunching calculation gave a methods for speeding the procedure of grouping by figuring the separation between the certainty proportions of the arrangement to the group head. These groups have evidently added to the quicker and precise order. The Naïve Bayesian characterization has been altered in the method of working by ascertaining the standard deviation measures and with the standard deviation accomplished the likelihood of the information having a place with a specific class has been accomplished. The factual measures have without a doubt demonstrated that the proposed calculations stand a path in front of all the current strategies.

REFERENCES

1. Heba Abusamra. "*A comparative study of feature selection and classification methods for gene expression data of glioma*", Procedia Science Direct, Elsevier Issue.10.1016/j.procs.2013.10.003. For Conference.
2. Jia Lv,Qinke Peng,Xiao Chen, Zhi Sun , "*A multi-objective heuristic algorithm for gene expression microarray data classification*", Elsevier, Expert Systems with Applications 59(2016)13-19.
3. Krisztian Buza, "*Classification of Gene Expression data: A Hubness-aware semi-supervised approach*", Elsevier, Computer Methods and Programs in Biomedicine 127(2016) 105-113.
4. Hung-Yi Lin, "*Gene Discretization based on EM clustering and adaptive sequential forward gene selection for molecular classification",* Elsevier, Applied Soft Computing 48(2016) 683-690..
5. Sara Tarek,Reda Abd Elwahab,Mahmoud Shoman, "*Gene Expression based cancer classification*", Egyptian Informatics Journal 2016.
6. Devi Arockia Vanitha ,Devaraj D,Venkatesulu, "*Gene Expression Data classification using support Vector Machine and Mutual Information–based Gene selection*", Procedia Computer science 47(2015)13-21.
7. Konstantina Kourou, Costas Papaloukas,Dimitrois I.Fotiadis, "*Integration of pathway Knowledge and Dynamic Bayesian Networks for the prediction of Oral Cancer Recurrence*", IEEE 2016.

8. W. Yip, K. Law, and W. Lee, "Forecasting Final/Class Yield Based on Fabrication Process E-Test and Sort Data," in *Proceedings of the IEEE International Conference on Automation Science and Engineering (CASE 07)*, Scottsdale AZ, USA, 2007, pp. 478-483.

9. S. Kotsiantis, D. Kanellopoulos, and P. Pintelas, "Handling imbalanced datasets: A review," *GESTS International Transactions on Computer Science and Engineering*, vol. 30, pp. 25-36, 2006.

10. T. M. Khoshgoftaar and K. Gao, "Feature Selection with Imbalanced Data for Software Defect Prediction," in *Proceedings of the Eighth International Conference on Machine Learning and Applications (ICMLA 09)*, Miami, Florida, USA, 2009, pp. 235-240.

11. P. Yang, L. Xu3, B. B. Zhou, Z. Zhang, and A. Y. Zomaya, "A particle swarm based hybrid system for imbalanced medical data sampling," *BMC Genomics Journal*, vol. 10, p. S34, 2009.

12.] P. Li, K. L. Chan, and W. Fang, "Hybrid Kernel Machine Ensemble for Imbalanced Data Sets," in *Proceedings of the 18th International Conference on Pattern Recognition (ICPR 06)*, Hong Kong, 2006, pp. 1108-1111.