RECURRENCE CAPTURE OF LIVER DISEASE USING SUPPORT VECTOR MACHINE WITH IMPROVED PARTICLE SWARM OPTIMIZATION

P. Laura juliet[1], T. Shanmugapriya[2]

[1] Assistant Professor, Dept. of Computer Applications, Vellalar College for Women, Erode, Tamilnadu, India
[2]Research Scholar, Department of Computer Science, Vellalar College for Women, Erode, Tamilnadu, India

---***---

Abstract - *The ability to discover patient acuity or severity of illness has immediate practical use for clinicians. Evaluate the use of multivariate time series modeling along with multiple models. To evaluate the data-merging algorithm, performance of prediction using processed multiple measurements are compared to prediction using single measurements. This scenario is used to perform the multiple time series data processing along with multiple measurements. The merging algorithm also statistical measures are performed and calculated based on the specified dataset. However it as issue with unbalanced data and classification performance is reduced significantly. To avoid this issue in proposed scenario, the proposed algorithm named as improved Particle swarm Optimization algorithm (IPSO) is used for feature selection. This algorithm is used to increase the prediction accuracy. The experimental result concludes that proposed system provides greater performance rather than existing scenario. SVM classifier is used to classify whether the disease is recurrence or not.*

Key Words: *Data Mining, Clinical Data, Liver Disease, Merging Algorithm, Classification, SVM, IPSO.*

1. INTRODUCTION

Data mining is the process of extracting patterns from data. It is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, statistics, machine learning, , and database systems. The overall objective of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Classification is a data mining (machine learning) techniques used to predict group membership for data instances. It is a data mining function that assigns items in a collection to target categories or classes. It involves finding the rules that partition the data into disjoint groups. Hepatocellular carcinoma (HCC), also called malignant hepatoma, it is the most common type of liver cancer. Most cases of HCC are as a result of either a viral hepatitis infection (hepatitis B or C), metabolic toxins such as alcohol or aflatoxin. Treatment options for HCC and prognosis are dependent on many factors but especially on staging, tumor size, and extent of liver injury. Signs and Symptoms of liver cancer: Small HCC produce no symptoms, Discovery of HCC when symptoms are present provides little value to the patient, Systemic symptoms of cancer such as anorexia, unintended weight loss, and local symptoms such as the right upper quadrant pain almost guarantee the disease is untreatable. Here by used HCC patient data set. SVM Classifier is used to classify HCC patient data whether to find out the diseases will be recurrence or not. Hence this IPSO algorithm improves the training process speed and reduces the time complexity of scenario. It is also used to increase the accuracy of classification results more effectively.

2. LITERATURE REVIEW

Yi-Ju Tseng, Xiao-Ou Ping [16], proposed a description of patient conditions should consist of the changes in and combination of clinical measures. Clinical data from 83 hepatocellular carcinoma (HCC) patients are used in this research. Their clinical reports from a defined period were merged using the proposed merging algorithm, and statistical measures were also calculated. After that the data processing, multiple measurements support vector machine (MMSVM) with radial basis function (RBF) kernels was used as a classification method to predict HCC recurrence. A multiple measurements random forest regression (MMRF) was used as, an additional evaluation/ classification method. The results of recurrence prediction by MMSVM with RBF using improved when the proposed data-processing algorithm was used, that multiple measurements could be of greater multiple measurements and a period of 120 days (accuracy 0.771) was optimal. The results show that the performance of HCC-recurrence prediction was significantly value than single. There were a total of 83 patients in the experiment dataset, of whom 18 were recurrent patients, and 65 were non recurrent patients.

Yan-Bo Lin [17], have proposed a methods of over-sampling and under-sampling are used for handling the issues of data imbalanced. The case based reasoning (CBR) is used for developing classification models to predict recurrent statuses of patients with liver cancer. According to the preliminary results of classification methods, on average, the BAC of balanced methods of the under-sampling (66.07%) and the over-sampling (54.24%). Most importantly, the under-sampling method could acquire the highest mean accuracy of the three datasets (under-sampling: 66.76%, over-sampling: 53.47%, imbalanced: 48.58%). Therefore, the balanced datasets could provide benefits for classification models and efficiently reduce biased interpretations.

YasunoriMinami and Masatoshi Kudo [18], has proposed Radiofrequency ablation (RFA) of liver cancers can be performed safely using percutaneous, laparoscopic, or open surgical techniques, and much of the impetus for the use of Radio Frequency Alabration(RFA) has come from cohort series that have provided an evidence base for this technique. An accurate evaluation of treatment response is very important to protected successful RFA therapy since a sufficient safety margin (at least 0.5 cm) can prevent local tumor recurrences. And also provide a profile of side effects and information on the integration of this method into the general management of patients with HCC. To minimize complications of RFA, physicians should be familiar with each feature of complication. Appropriate management of complications is essential for successful RFA treatment. The use of a laparoscopic or open approach allows repeated placement of RFA electrodes at multiple sites to ablate larger tumors. In addition, an accurate valuation of treatment response is very important to secure successful RFA therapy since a sufficient safety margin (at least 0.5 cm) could prevent local tumor recurrence. Adjuvant therapy, such as molecular targeted therapies following curative therapy, is expected to further improve survival after RFA.

Xiao-Ou Ping, MS, Yi-Ju Tseng [5], has proposed an efficient way for tracking patients condition over long periods of time and to facilitate the collection of clinical data from different types of narrative reports, it is critical to develop an efficient technique for smoothly analyzing the clinical data accumulated in narrative reports. The classifier provided the answers and direct/indirect evidence (evidence sentences) for the clinical questions. The performance of the rule-based classifier achieved an accuracy from 96.15% to 100%, PPV from 94.12% to 100%, NPV from 82.35% to 100%, sensitivity from 95.31%to 100%, and specificity from 95.56% to 100% .

Zahra Beheshti, Siti MariyamHj [8], analyzed an Enhancement of artificial neural network learning using centripetal accelerated particle swarm optimization for medical diseases diagnosis. ANN is also used in the medical diseases diagnosis. To evolve the ANN learning and accuracy, a new meta-heuristic algorithm, centripetal accelerated particle swarm optimization (CAPSO) is applied. The hybrid learning of CAPSO and multi-layer perceptron (MLP) network are used to classify the data. The efficiency of the methods is evaluated based on mean square error, accuracy, precision, recall, and area under the receiver operating characteristics (ROC) curve. The result shows that this method gives better performance than other methods in terms of testing data and data sets with high missing values.

2.1 Problem Definition

The multiple time series using random forest has still issue with optimal accuracy in classification results. The existing scenario has time complexity for various datasets. It also has problem with unbalanced dataset. Thus the system performance is reduced prominently.

2.2 Existing System

This system analyzes the clinical dataset by using the feature selection, data reduction, and extraction and transformation techniques. Hereby, introduced the scheme named as multiple measurements support vector machine (MMSVM) and multiple measurements random forest (MMRF)for predicting results. Initially it is to be considered as the single measurement data and multiple measurement data. Then apply the feature selection and perform the training as well as testing process. Then the selection of features is combined by using merging algorithm for time series dataset. Also by using merging algorithm we achieve the multiple time series with multiple measurements. The dataset are, LIS laboratory information system, RIS: radiology information system, HIS: hospital Information systems. This existing algorithm obtains more useful information for HCC recurrence prediction. After that, the collected dataset multiple features are merged using algorithm with various time sequences. Then it has to compute the statistical measure and important features are selected from the specified dataset. The particular model is generated and based on this model that can be classifying the predicted results.

2.3 Proposed System

To overcome the above mentioned issues this project goes for proposed scenario by using optimization algorithm. The proposed scenario introduced the algorithm named as improved particle swarm optimization (IPSO) is also called as modified multiple swarm PSO (MSPSO). This optimization algorithm is focused on the improvement of more optimal feature selection on different time series dataset. The Particle Swarm Optimization (PSO) is a Meta - heuristic search technique, biologically inspired from the nature's social behavior, dynamic movements and communications of insects, birds and fish. It is a population based stochastic global optimization technique evolved to study the social behavior of insects, birds or fish as why they move in a group searching for food randomly in some area, knowing only the distance from the food. A PSO system combines local search methods (through self-experience) with global search methods (through neighboring experience), attempting to balance exploration and exploitation. PSO has been successfully applied to various areas including Prediction analysis. The unbalanced dataset are balanced and misclassification results are eliminated in this scenario. The IPSO algorithm targets on the reduction of unrelated feature data and progress of more relevant data for time series data. Hence this algorithm improves the training process speed and reduces the time complexity of scenario. It is also used to increase the accuracy of classification results more effectively. SVM Classifier is used to classify HCC patient data whether to find out the diseases will be recurrence or not. Thus the problem of unbalanced dataset is handled efficiently in this scenario by using the improved PSO algorithm. From the result the proposed system yields more accurate classification results for various time period datasets.

Advantages of the proposed system

- It increases the classification accuracy in superior.
- It reduces the computational complexity.
- It selects the more optimal features.
- It improves the system performance in higher.

3. SYSTEM METHODOLOGY

Detect whether the liver cancer HCC recurrence from LIS laboratory information system databases.

- Select the HCC (Heptocellular carcinoma) liver cancer dataset and analyze data.
- The feature selection will be done using IPSO (Improved particle swam optimization) Algorithm.
- Apply the data in two measurements,
  - ➢ single measurement
  - ➢ multiple measurement
- By doing this measurement test, it detect whether the HCC recurrence in which of these two measurements.
- Single & multiple measurement performance is based on weights and time periods. Based on the performance analysis result is obtained using their respective ways.
- By using merging algorithm multiple time periods will be merged.
- Evaluate the Statistical calculation for getting the result.
- Then k-fold technique is applied to the obtained result.
- Then the data will be segregate into equal parts for training and testing data.
- Apply SVM to classify the data.
- By using the SVM classifier classify the obtained result.
- The predicted value is generated which is better when comparing single measurement data than multiple measurement data.
- Accuracy will be analyzed.
- Finally, performance metrics was evaluated.

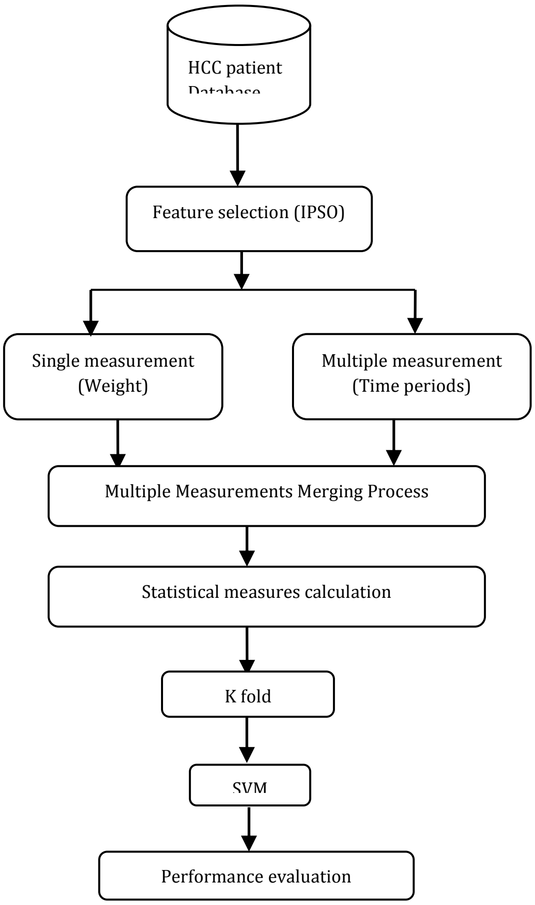

HCC patient Database
Feature selection (IPSO)
Single measurement (Weight)
Multiple measurement (Time periods)
Multiple Measurements Merging Process
Statistical measures calculation
K fold
SVM
Performance evaluation

Fig-1: System flow

List of Modules

- Preprocessing
- Feature selection(IPSO)
- Merging Algorithm Based on Defined Time Periods
- Calculation of statistical measure
- Prediction model establishment
- SVM

3.1 Preprocessing

In this module, the pre processing technique is performed to obtain the more accurate classification results. Data cleaning is the process of discovering and correcting inaccurate records from the specified dataset. Used mainly in databases, the term refers to identifying incomplete, incorrect, inaccurate, irrelevant (Noise reduction) etc. It is used to increase the classification accuracy results for the specific query By using preprocessing method, as we all know that real world data contains missing values or noisy values so in order to produce good results from the data set we need to mine data.

3.2 Feature selection

In this module we have to perform the feature selection process on the time period dataset. It is used to provide relevant feature to the training and testing process. To allow us to compare the performance of prediction by adding statistical measures with others, it set the maximum number of six feature values such as minimum, maximum, standard deviation, and variant, mean, conventional. Combinations of these features were then created by sequentially adding features, and were input into classification algorithms to generate the prediction model. IPSO algorithm is used for selecting the features. The algorithm proposed in this work is based on the particle swarm optimization technique. PSO is a computational algorithm that iteratively optimizes a given solutions by applying mathematical rules and after estimate the fitness of a current solutions changes their coordinates into the search space. PSO utilizes a certain number of solutions, called particles that form a swarm. Every such particle has position and velocity coordinates in the search space. The velocity represents the change of the particle position from iteration to iteration.

v[ ] = v[ ] + c1 * rand() * (pbest[ ] - present[ ]) + c2 * rand() * (gbest[ ] - present[ ]) (a) Present [ ] = present [ ] +v [ ](b)

V [ ] is the particle velocity,
Present [ ] is the current particle (solution),
pbest [ ] and gbest [ ] are defined as position and global Best,
rand ( ) is a random number.

The change of the particle's position is dictated by the best so far known particle's position as well from the best position in the overall swarm. This is used to improve the speed of the process by using important and relevant information features in the dataset. It reduces the number of iterations by selecting the best solutions. The IPSO algorithm is as follows:

```
Update PSO
{
Do
ForEach Particle in Swarm
For j = 0 to ParticleLength
Partcle.Velocity[j] = W * Partcle.Velocity[j] + C1*R1*Particle.BestPosition[j] - Particle.Position[j] +C2*R2*BestParticle.Position[j] - Particle.Position[j]
EndFor
For j = 0 to ParticleLength
Partcle.Position[j] += Partcle.Velocity[j]
EndFor
CheckCandidate (Particle)
If (Particle.BestInfoGain > BestParticle.BestInfoGain)
BestParticle = Particle
EndIf
EndForEach
OldBestGain = NewBestGain
NewBestGain = GetSwarmBestInformationGain
While ( (OldBestGain - NewBestGain) > EPSILON )
BestShapelet = BestParticle
}
```

3.3 Merging Algorithm Based on Defined Time Periods

This module has to merge the important statistical values by using the merging Algorithm more efficiently. Based on this merging algorithm to evaluate the time periods data. The central idea of this merging algorithm is to choose only one value to stand for a feature in one period. Because the time of target event, therapy for HCC those are set as the key time with regard to data processing. The value that is nearer to event time could be more significant than others. Therefore, the most recent value is selected to represent a feature in a period, and some valuable information in the original data might be omitted by the merging Algorithm.

3.4 Calculation of statistical measure

Statistical measures were calculated for describing the data circulation in each period. There was a possibility that information in the original data, such as the tendency and the distribution of the features, may disappear after data merging. To preserve this information, the definitions of statistical measures are considered that are shown. The maximum and minimum are used for illustrating the extremes of the data. Average is a method for deriving the middle tendency of a sample space, and standard deviation is a widely used measurement of variability or diversity. Pearson's correlation coefficient is a statistical technique, shows whether and how strongly pairs of features are related, and express this relationship in values ranging between −1 and 1. The closer, complete value of the Pearson's coefficient is to 1, the stronger the correlation between the features is. A trend line represents the long-term movement in time-series data, and it tells whether a particular measure have increased or decreased in over the period of time. These statistical measures were used for expressing the data distribution in a exact period, and the information lost during data merging might be partly retained in them.

3.5 Prediction model Establishment

This project used data mining algorithms for single-measurement data and also multiple-measurements data. The SVM is a data-mining method that constructs a prediction model for a binary class. It uses nonlinear mapping to convert the data into a higher dimension. With an appropriate nonlinear mapping to a sufficiently high dimension, data from two classes are separated by a hyper plane. The LIB SVM library is used for implementing the SVM classification method. The most popular kernel function, RBF, was used for SVM model establishment. For multiple measurements, the classification results are decided by a voting mechanism: where one or more instances belonged to the same group (patient), the class with the most votes in these instances was the final result of classification. The MMSVM is a tool to developed and enable LIB SVM to conduct cross validation and prediction with multiple measurements' results from the voting mechanism.

3.6 SVM Classifier

In machine learning, Support Vector machines(SVM) are supervised learning models with connected learning algorithms that analyze data used for classification An SVM classifier classifies the data by determining the optimal Hyper plane which can separates all the data points of one class from that of the other class. This optimal hyper plane for an SVM classifier means the one with the highest margin between the two classes. The margin means the maximal width of the slab parallel to the hyper plane that has no interior data points. SVM classifier is used to classify the patient data based on the class label (0,1). When a class label generated the value 0 then it meant to be non recurrence of diseases. Otherwise if class label generate 1 then it meant to be recurrence of diseases.

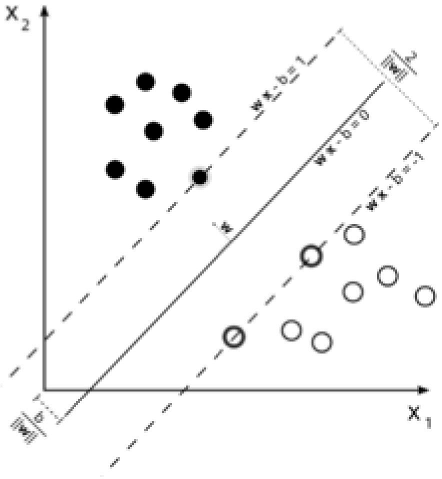


Fig- 4.2 SVM Classifier

Maximum-margin hyper plane and margins for an SVM trained with samples from two classes. Sample on the margin are called the support vectors.

4. RESULTS AND DISCUSSIONS

To execute the proposed technique and generate various result using Mat Lab Tool in this environment. The scenario is about HCC patient dataset to discover the HCC recurrence prediction. The analysis has been done for existing and proposed research work by using IPSO algorithm. The performance metrics are such as accuracy, precision, recall and F-measure from the investigational result the conclusion decides that the proposed method provides higher performance result in terms of greater accuracy.

Dataset

The HCC recurrence was predicted through ipso algorithm then the medical team member carefully analysis relevant literature .According to the classification scheme a particular disease was detected for the HCC patient dataset are follows. Dataset consist of nine attributes and 607 instances. Eight attribute values are Numerical. The attributes are ALP.B, ALT.B, AST.B, TBL.B, ALP.M, ALT.M, AST.M, TBL.M (numerical).And the one attribute is DOSE (character).

Accuracy

The accuracy percentage of true results (both true positives and true negatives) among the total number of cases examines. Accuracy refers to the proximity of an exact value to a standard or known value.

Accuracy can be calculated from formula given as follows

$$Accuracy = \frac{True\ positive + True\ negative}{True\ positive + True\ negative + False\ positive + False\ negative}$$

Precision

Precision value is evaluated according to the feature classification at true positive false positive prediction. Precision is a description of random errors, a measure of statistical variability. It is expressed as follows

$$Precision = \frac{True\ positive}{True\ positive + False\ positive}$$

Recall

Recall value is evaluated according to the feature classification at true positive prediction, false negative. *Recall* in memory refers to the mental process of retrieval of information from the past. It is given as,

$$Recall = \frac{Truepositive}{(Truepositive + Falsenegative)}$$

F-Measure

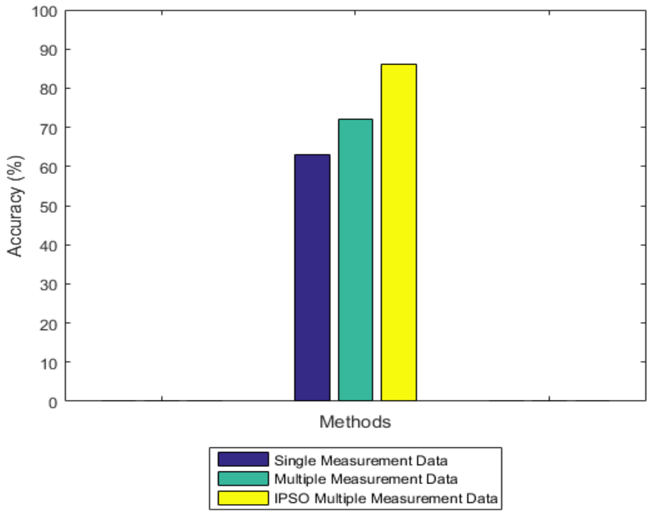
F-measure is calculated from the precision and recall value. Precision is also used with recall, the percent of all relevant documents that is return by the search. The two measures are sometimes used together in the F1 Score to provide a single measurement for a system. It is calculated as,

F-measure = $2 \times (precision \times recall / precision + recall)$

4.2 Comparison Result

Table-1: Comparison table

Metrics	Single Measurement Data	Multiple Measurement Data	IPSO Multiple measurement data
Accuracy	63	72	87
Precision	0.6341	0.6989	0.8407
Recall	0.6595	0.7333	0.8690
F-Measure	0.6465	0.7157	0.8546



Chat-1: Accuracy

5. CONCLUSION AND FUTURE WORK

This section concludes that proposed scenario is better by using the improved particle swarm optimization algorithm. The different time periods data is evaluated and the methodology is focused on the prediction of more accurate results. Existing system presents a merging algorithm for multiple-time series data with different sampling rates and data types, and calculates the consequence of adding statistical measures to it. The results show that the performance of HCC-recurrence prediction is significantly improved through use of the algorithm, and as a corollary, that multiple measurements may provide more useful information for HCC-recurrence prediction than single measurement does. However it has issue with the classification results in terms inaccuracy due to unbalanced dataset. To avoid this issue in proposed scenario enhanced the algorithm named as improved particle swarm optimization algorithm. The IPSO is focused on the improvements of optimization of classification performance. It reduces the overall training time computation complexity and also it allows more relevant and optimal solution for the classification process. Thus it achieved higher prediction accuracy than existing system. The result can prove that the proposed system is better to existing system by using IPSO algorithm. This research can be extended in future with following scopes:

- Apply Inverse random sampling techniques that can be used in future for facing the imbalance problem in the dataset.
- Invitation to explore the potential of other machine learning techniques.

REFERENCES

[1] Ai-Qin Mu1,2, De-Xin Cao1,“A Modified Particle Swarm Optimization Algorithm”, Natural Science Vol.1, No.2, 151-155 (2009).

[2] Alex S. Befeler And Adrian M. Di Bisceglie, “Hepatocellular Carcinoma:Diagnosis And Treatement”, Gastroenterology 2002.

[3] Ashis Pradhan” SUPPORT VECTOR MACHINE-A Survey” International Journal of Emerging Technology and Advanced Engineering, (ISSN 2250-2459, Volume 2, Issue 8, August 2012).

[4] Gopala Krishna Murthy Nookala, Bharath Kumar Pottumuthu,“ Performance Analysis and Evaluation of Different Data Mining Algorithms used for Cancer Classification” , (IJARAI) International Journal of Advanced Research in Artificial Intelligence, Vol. 2, No.5, 2013.

[5] Guan-Tarn Huang, Xiao-Ou Ping,MS, Yi-Ju Tseng, “Original Research Information Extraction For Tracking Liver Cancer Patients’ Statuses: From Mixture Of Clinical Narrative Report Types, DOI: 10.1089/Tmj.2012.0241 Mary Ann Liebert, Inc. _ Vol. 19 No. 9 _ September 2013.

[6] Hany M. Harb, “Feature Selection on Classification of Medical Datasets based on Particle Swarm Optimization” , International Journal of Computer Applications (0975 – 8887) Volume 104 – No.5, October 2014.

[7] Neha, "Particle Swarm Optimization based Feature Selection" , International Journal of Computer Applications (0975 – 8887) Volume 146 – No.6, July 2016.

[8] Priyangaand Dr, Prakasam.S " The Role of Data Mining-Based Cancer Prediction system (DMBCPS) in Cancer Awareness"International Journal of Computer Science and Engineering Communications- IJCSEC. Vol.1 Issue.1, December 2013.

[9] Reetu,"Medical Diagnosis for Liver Cancer using Classification Techniques", International Journal of Science and Research (IJSR) Volume 4 Issue 5, May 2015.

[10] Ryosuke Tateishi And Shuichiro Shiina," Prediction of Recurrence of Hepatocellular Carcinoma After Curative Ablation Using Three Tumor Markers", Hepatology, Vol. 44, No. 6, 2006.

[11] Rajeswari. P, "Human Liver Cancer Classification using Microarray Gene Expression Data", International Journal of Computer Applications (0975 – 8887) Volume34– No.6, November 2011.

[12] Shomona Gracia Jacob" Data Mining in Clinical Data Sets: A Review" International Journal of Applied Information Systems (IJAIS) – ISSN: 2249-0868.

[13] ShwetaKharya, "Using Data Mining Techniques for Diagnosis of Cancer Disease", International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.2, April 2012

[14] Stacey.M and McGregor. C, "Temporal abstraction in intelligent clinical data analysis: A survey," Artif Intell Med, vol. 39, no. 1, pp. 1–24, 2007.

[15] Tiago Sousa , Arlindo Silva, " Particle Swarm based Data Mining Algorithms for classification tasks",(ELSEVIER) T. Sousa et al. / Parallel Computing 30 (2004) 767–783.

[16] Yi-JuTseng,Xiao-OuPing,"Multiple-Time-Series Clinical Data Processing For Classification with Merging Algorithm and Statistical Measures," IEEE Journal of Biomedical and Health Informatics, Vol. 19, No. 3, May 2015.

[17] Yan-Bo Lin I, Xiao-Ou Pint, "Processing and analysis of imbalanced liver cancer Patient data by case-based reasoning", The 2014 Biomedical Engineering International Conference (Bmeicon-2014).

[18] YasunoriMinami and MasatoshiKudo, "Radiofrequency Ablation of Hepatocellular Carcinoma: A Literature Review", International Journal of Hepatology Volume 2011, Article ID 104685, 9 pages doi:10.4061/2011/104685.