International Journal of Engineering & Technology, 7 (3) (2018) 1114-1118

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET
doi: 10.14419/ijet.v7i3.10693
Research paper

Predicting hyperlipidemia using enhanced ensemble classifier

Lakshmi K.S.[1]*, G. Vadivu[2], Suja Subramanian[3]

[1]*Assistant Professor, Department of Information Technology, Rajagiri School of Engineering & Technology, Kochi, Kerala, India*
[2]*Professor, Department of Information Technology, SRM University, Kattankulathur, Chennai, Tamil Nadu, India*
[3]*Research Scholar, Department of Electronics, Cochin University of Science & Technology, Kerala, India*
**Corresponding author E-mail: lakshmiks@rajagiritech.edu.in*

Abstract

Advancement in medical technology has resulted in bulk creation of electronic medical health records. These health records contain valuable data which are not fully utilized. Efficient usage of data mining techniques helps in discovering potentially relevant facts from medical records. Classification plays an important role in disease prediction. In this paper we developed a prediction model for predicting hyperlipidemia based on ensemble classification. Support Vector Machine, Naïve Bayes Classifier, KNN Classifier and Decision Tree method are combined for developing the ensemble classifier. Performance of each classifier is evaluated separately. An overall accuracy of 97.07% has been obtained by using ensemble approach which is better than the performance of each classifier.

Keywords: *Classification; Decision Tree; Hyperlipidemia; Ensemble classifier; Naïve Bayes; Support Vector Machine.*

1. Introduction

Data Mining plays a vital role in extracting useful information from huge amount of data. Some of the major application areas of data mining include banking, health care, business, marketing, and transportation. Classification, Frequent Pattern Mining, Association Rule Mining and Clustering are the major techniques employed in data mining for discovering novel facts from massive data.

Recent advancements in technology have leveraged the production of huge volumes of data in health care domain. At whatever point a patient visits a doctor or other medicinal supplier, a record is kept known as medical health record (MHR). Most of them have an institutionalized configuration and are mostly paper-based. Manual processing of these archives for information extraction is costly and can cause mistakes. Therapeutic records contain a lot of important data for directing looks into. For better patient care, powerful administration of therapeutic data is vital. The accessibility of up-to-date disease profiles might be important for an assortment of applications including suggesting medications, replying clinical inquiries and speculation disclosure.

Medical health records can be a clinical note, release synopsis or agent report. Data related with various illnesses can be extricated and incorporated from these content hotspots for better administration of diseases. There are numerous systems available for disease prediction. Major disease domains were data mining techniques are used for prediction include heart disease [1], [2], [3], [4], kidney disease [5], [6] diabetes [7], [8] and liver disease [9], [10], [11].

2. Metabolic Syndrome

Metabolic syndrome is not a disease itself. It is a group of risk factors which includes increased blood pressure, high blood sugar, excess body fat around the waist, and abnormal cholesterol or triglyceride levels that occur together, increasing the risk of heart disease, stroke and diabetes [12], [13]. These risk factors double risk of blood vessel and heart disease, which can lead to heart attacks and strokes. They increase risk of diabetes by five times. Metabolic syndromes are closely associated with cholesterol. Variations in the level of cholesterol can lead to metabolic syndrome. Metabolic syndrome is becoming more common nowadays. But the good news is that it can be controlled, largely with changes to your lifestyle. Risk factors [14] of metabolic syndrome are shown in Table 1. Prediction of metabolic syndrome can be done based on the analysis of these risk factors. In [15], authors developed a system for metabolic syndrome prediction using random forest method.

Table 1: Metabolic Syndrome Risk Factors

Risk Factors	Men	Women
Large Waist Size	≥ 40”	≥ 35”
High Triglycerides	≥ 150mg/dL	≥ 150mg/dL
Low HDL	< 40mg/dl	< 50mg/dl
High BP	≥ 135/85mmHg	≥ 135/85mmHg
Fasting Blood Sugar	≥ 100mg/dL	≥ 100mg/dL

3. Hyperlipidemia

Cholesterol is a fat-like substance that is present in all human body cells. It helps in the production of vitamins, hormones and digestive enzymes thereby enabling our body to work properly. Cholesterol travels through bloodstream in the form of lipoproteins. There are two kinds of lipoproteins: LDL (Low Density Lipoproteins) and HDL (High Density Lipoproteins). LDL is considered as bad cholesterol and HDL as good cholesterol. Both these cholesterol are needed in healthy levels for proper functioning of the body. The medical term for high blood cholesterol is lipid disorder, or Hyperlipidemia. Hyperlipidemia is diagnosed using lipid profile [16]. Significant parameters in Lipid Profile are the following:

- Total Cholesterol(TC)
- High-Density Lipoprotein Cholesterol(HDL)
- Low-Density Lipoprotein Cholesterol(LDL)

- Very-Low-Density Lipoprotein (VLDL)
- Triglycerides(TG)

3.1. Triglycerides

Triglycerides (TG) are a type of fat found in our blood. Triglycerides in low amount are needed for our good health. But high level of triglycerides can increase risk of heart disease and metabolic syndrome. TG is used to measure lipid status and metabolic disorders. They are the major component of chylomicrons and VLDL, two types of lipoproteins. They may be elevated in diabetes, hypothyroidism, chronic liver and kidney diseases, pancreatitis, some genetic types of hyperlipidemia, alcohol abuse, estrogen (pregnancy or oral contraceptive pills), and certain medications (thiazide diuretics). A patient must absolutely be fasting for an accurate measurement. Reference range of triglycerides [17] is shown in Table 2.

Table 2: Reference Range For Triglycerides

Unit (mg / dL)	Range
< 150	normal
150 – 199	Borderline-high
200 – 499	high
> 500	very high

3.2. HDL (High Density Lipoprotein)

HDL or High Density Lipoprotein is called the good cholesterol. It tends to carry bad cholesterol away from tissues. People with high HDL are at lower risk for heart disease. People with low HDL are at higher risk. Reference range of HDL is shown in Table 3.

Table 3: Reference Range For HDL

Unit (mg / dL)	Range
< 40	low
> 60	high

3.3. LDL (Low Density Lipoprotein)

LDL cholesterol is called the bad cholesterol. It is part of the lipid profile and is one of the more important risk factors for atherosclerotic (CHD) disease. LDL is the cholesterol component that binds to liver receptors and tends to control the formation of cholesterol. The Friedewald formula (FF) is the main method for evaluating low-density lipoprotein cholesterol (LDL-c). The formula can only be used when the TG are less than 400 mg/dL. LDL core lipids contain about 10% TG and 45% cholesterol. Table 4 shows the reference range of LDL.

Table 4: Reference Range For LDL

Unit (mg / dL)	Range
< 100	Optimal
100 - 129	Near optimal
130 - 159	Borderline high
160 - 189	High
> 190	Very high

4. Data Set

The data set used in this project is clinical data set collected from hospitals around Cochin and contain records of around 700 patients. The clinical data set specification provides succinct, explicit definition for items related to hyperlipidemia and diabetes. Biochemistry results are collected from the hospital database. Data set include the attributes mentioned in Table 5.

Table 5: Dataset Description

SI. No.	Attributes	Descriptions
1	Sex	Patient gender
2	Age	Patient age
3	FBS	Fasting blood sugar
4	TC	Total Cholesterol
5	TG	Triglycerides
6	HDL	High Density Lipoprotein
7	LDL	Low Density Lipoprotein
8	VLDL	Very Low Density Lipoprotein
9	HL?	Hyperlipidemia
10	MB?	Metabolic Syndrome

5. Data Mining

Data Mining is the extraction of hidden facts from massive databases. There are various techniques used in data mining [18]. Classification, Clustering and Association Rule Mining are the most widely used data mining techniques. In this project, we have used classification, which is a supervised learning [19] method.

5.1. Classification

Classification is a data mining process by which each item is assigned to a particular class based on the features associated with the item under consideration.

Classification Algorithms can be broadly classified into 3 types:

1. Based on Frequency Table
2. Based on Covariance Matrix
3. Based on Similarity Functions

Frequency table based classification algorithms are ZeroR, OneR, Naïve Bayes algorithm and Decision Tree Induction algorithm. Algorithms based on Covariance Matrix are Logistic Regression and Linear Discriminant Analysis (LDA). K-Nearest Neighbor (kNN) classifier is based on similarity functions. Apart from these algorithms, there are algorithms like Support Vector Machine (SVM) and Artificial Neural Network (ANN) used for classification. In this paper, we have used an ensemble approach for classification.

5.2. Ensemble Classifier

In the ensemble approach, various classification techniques are combined in order to improve the accuracy of the entire model. Various researches have been conducted for developing efficient systems using ensemble approach [20], [21], [22], [23]. In this project, we have used SVM, Naïve Bayes, KNN and Decision Tree for developing the ensemble classifier.

Support Vector Machine is a supervised classification learning algorithm. It is based on the concept of decision planes that define decision boundaries. A decision plane is one that separates between a set of objects having different class memberships. SVM classifier finds the best hyperplane for separating data points that belong to a particular class from the data points of other classes. The best hyperplane is the hyperplane that has the largest margin between the two classes. Margin means the maximal width of the slab parallel to the hyperplane that has no interior data points. The support vectors are the data points that are closest to the separating hyperplane; these points are on the boundary of the slab.

Features of SVM Classifier are:

- Accuracy is high
- Less over-fitting

- Robust to noise
- Not trapped in local minima
- Works well with less training examples

There are several prediction systems developed using SVM classifier [24], [25], [26].

Naive Bayes classifier is based on Bayes theorem. It uses strong (naive) independence assumptions between the features, which presume that an attribute value on a given class is independent of the values of other attributes.
Features of Naïve Bayes Classifier are:

- Simple to implement
- Not sensitive to noisy data
- Can handle both real and discrete data
- Great computational efficiency
- Accurate for most of the classification

Many systems have been developed using Naïve Bayes classifier for heart disease prediction [2], [27], [28]. Naïve Bayes classifier is a widely used classifier for many other disease prediction systems. [29], [30].

K-nearest neighbor's algorithm (k-NN) is a non-parametric method used for classification. Input consists of *k* closest training examples in the feature space. The output is a class membership. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its *k* nearest neighbors (*k* is a positive integer, typically small). If $k = 1$, then the object is simply assigned to the class of that single nearest neighbor.
Features of kNN Classifier are:

- It is based on a similarity function.
- No assumptions about the characteristics of the concepts to learn have to be done
- Complex concepts can be learned by local approximation using simple procedures
- Very simple classifier that works well on basic recognition problems.

Decision Trees (DTs) are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. A decision tree is a graph that uses a branching method to illustrate every possible outcome of a decision. Each node of the tree denotes an attribute.
Features of Decision Tree are:

- Implicitly perform variable screening or feature selection
- Require relatively little effort from users for data preparation
- Nonlinear relationships between parameters do not affect tree performance
- Easy to interpret

6. Proposed System

The overall architecture of the system is given in Figure 1. The proposed system is divided into 2 phases:

1. Ensemble Classification
2. Association Rule Generation

The proposed system can perform two mining techniques. Both predictive and descriptive models are used in this approach. Predictive models are used for Hyperlipidemia and Metabolic Syndrome Classification. Descriptive models are used for finding strong associations from the input dataset. In order to design such a system we have chosen a data mining tool suited to the medical diagnosis application for making highly accurate predictions. Classification task is accomplished using ensemble approach.

There are mainly 3 phases in classification process:

1. Sampling
2. Classification
3. Prediction

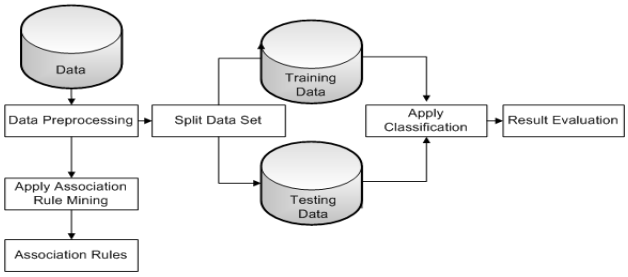

Fig.1: Overall System Architecture

In the first phase, the entire dataset is divided into 4 sets. Each dataset is then fed as input to the classifiers. A set of powerful machine learning algorithms are chosen for learning purpose. They are capable of predicting new observations from other observations after executing a process called learning from existing data. The algorithms try to fit a model closest to the characteristics of data under consideration. Models can be predictive or descriptive. Predictive models are used to make predictions. Descriptive models are used to identify patterns in data. Classification, regression, and time series analysis are some of the tasks of predictive modeling. Figure 2 shows training phase.

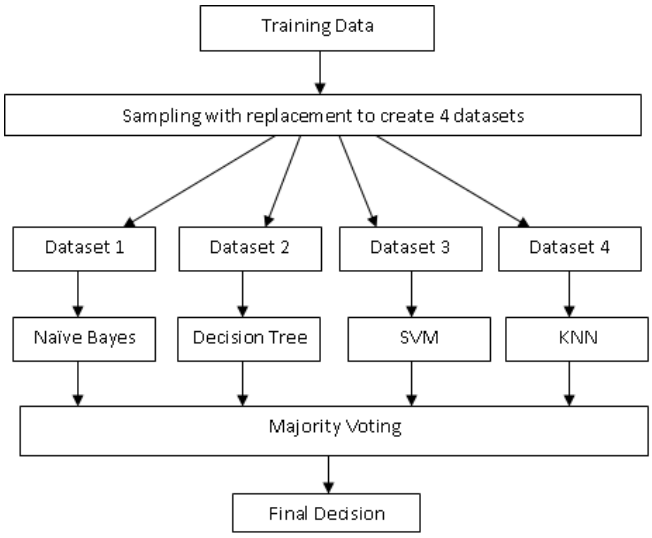

Fig.2: Training Phase

In this project unsupervised learning scheme is used, so after preprocessing the dataset contain an expected target field. This field is not considered for testing the developed model. Sampling is done on the input dataset and is fed as input to each of the classifier. The output of each classifier is then analyzed for measuring the accuracy of each classifier. Sampling with replacement is done so as to perform four iterations of training for each classifier. The average accuracy obtained from four epochs of each classifier is considered as the effective accuracy of that particular classifier. This accuracy is later considered for resolving tie in voting stage. For testing also, the dataset is divided into 4 samples. The detailed algorithm is given below:

Proposed Algorithm:
Input: Dataset divided into training set (Tr) and test data (Ts)
Step1: Set j=1
Step2: Set a counter for each of the classifier denoted as c_i initialized to 0.
Step3: Four random samples are created from Tr with replacement

Step4: Each sample dataset of Tr is fed as input to the classifier
Step5: Test the classifier with input Ts
Step6:If output of classifier$_i$ is positive increment c_i
Step7: Shuffle the sample dataset of Tr among each classifier and increment j
Step8: Repeat steps 2 to 8 till j=4
Step9: Set vote=0
Step10: If c_i>2,
then set output of classifier$_i$ as positive
otherwise negative
Step11: Increment the vote for each positive prediction
Step12: If vote>2 then the resultant class is positive
Elseif vote<2 then the resultant class is negative
else
choose the output with the highest accuracy

Table 6: Evaluation of various classifiers

Parameter	Naïve Bayes	Decision Tree	SVM	KNN
Accuracy	88.89	86.54	93.56	84.79
Sensitivity	91.79	89.55	96.26	88.05
Specificity	78.37	75.67	83.78	72.97
Precision	93.89	93.02	95.56	92.18
F-Score	92.82	91.25	95.90	90.06

A class is considered as positive, if the patient has hyperlipidemia otherwise the class is treated as negative.

7. Results

Performance of classifiers is measured mainly on the basis of following parameters:

a) Accuracy
b) Sensitivity
c) Specificity
d) Precision
e) F-measure

All these measurements are done using four quantities:

- True Positive (TP) – Number of hyperlipidemic patients who were classified as hyperlipidemic.
- True Negative (TN) – Number of people who are not hyperlipidemic patients and rightly classified as not hyperlipidemic.
- False Positive (FP) – Number of patients who were misclassified as hyperlipidemic but actually not hyperlipidemic.
- False Negative (FN) – Number of patients who were misclassified as not hyperlipidemic but actually they were.

Accuracy is the most instinctive performance measure and it is simply the ratio of correctly predicted observation to the total observations.

$$Accuracy = (TP + TN)/(TP+TN+FP+FN)$$

Sensitivity (Recall) is a statistical measure of how well a binary classification test correctly identifies a condition.

$$Sensitivity = TP/(TP + FN)$$

Specificity is a statistical measure of how well a binary classification test correctly identifies the negative cases.

$$Specificity = TN/(TN + FP)$$

Precision - Precision is the ratio of correctly predicted positive observations to the total predicted positive observations.

$$Precision = TP/TP+FP$$

F1 score - F1 Score is the weighted average of Precision and Recall.

$$F1\ Score = 2*(Recall * Precision) / (Recall + Precision)$$

Table 6 shows the evaluation of various classifiers. Performance comparison of classifiers is shown in Figure 3.

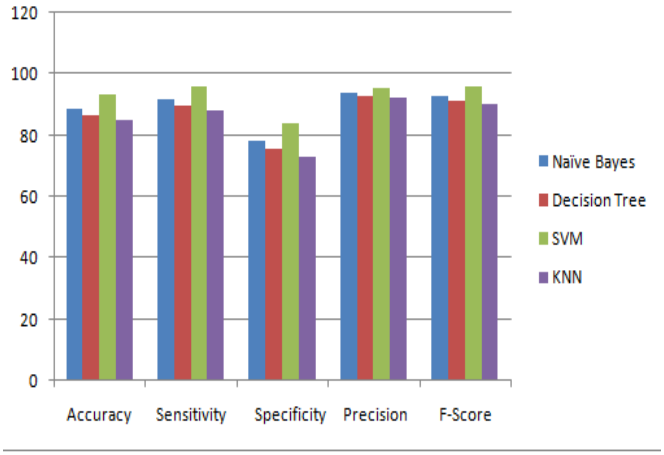

Fig.3: Performance comparison of classifiers

After analyzing the performance of individual classifiers on independent sets of training data, the enhanced classifier performance was computed based on majority voting. Value of various performance measures of the enhanced classifier is given in Table 7 and the comparison is shown in Fig.4.

Table 7: Performance of ensemble classifier

Parameter	Enhanced classifier
Accuracy	97.07
Sensitivity	98.5
Specificity	91.89
Precision	97.78
F-Score	98.13

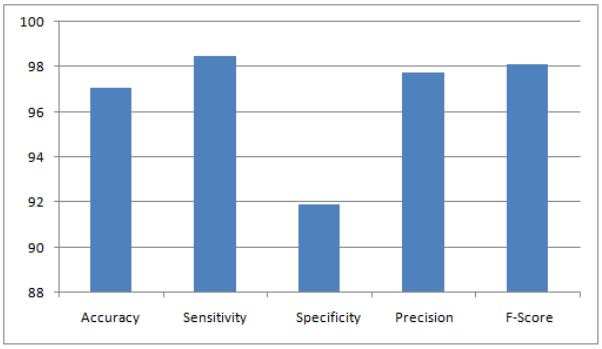

Fig.4: Performance comparison of ensemble classifier

References

[1] Hlaudi Daniel Masethe, Mosima Anna Masethe, "Prediction of Heart Disease using Classification Algorithms," Proceedings of the World Congress on Engineering and Computer Science 2014 Vol II WCECS 2014, 22-24 October, 2014.
[2] Shadab Adam Pattekari and AsmaParveen, "Prediction System For Heart Disease Using Naive Bayes," International Journal of Advanced Computer and Mathematical Sciences ISSN 2230-9624. Vol 3, Issue 3, 2012, pp 290-294.

[3] Purushottam, Kanak Saxena, Richa Sharma, "Efficient Heart Disease Prediction System", Procedia Computer Science, vol. 85, pp. 962, 2016, ISSN 18770509.

[4] Swathi P, Yogish HK, Sreeraj RS. "Predictive data mining procedures for the prediction of coronary artery disease", International Journal of Emerging Technology and Advanced Engineering. 2015, 5(2):339–42.

[5] S. Vijayarani1 ,S.Dhayanand, "Data Mining Classification Algorithms for Kidney Disease Prediction," International Journal on Cybernetics & Informatics (IJCI) Vol. 4, No. 4, August 2015.

[6] JamshidNorouzi, Ali Yadollahpour, Seyed Ahmad Mirbagheri, MitraMahdaviMazdeh, and Seyed Ahmad Hosseini, "Predicting Renal Failure Progression in Chronic Kidney Disease Using Integrated Intelligent Fuzzy Expert System," Computational and Mathematical Methods in Medicine, vol. 2016, Article ID 6080814, 9 pages, 2016. https://doi.org/10.1155/2016/6080814.

[7] AiswaryaIyer, S. Jeyalatha and RonakSumbaly, "Diagnosis Of Diabetes Using Classification Mining Techniques," International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.5, No.1, January 2015.

[8] HanWu, ShengqiYang, ZhangqinHuang, JianHe, Xiaoyi Wang, "Type 2 diabetes mellitus prediction model based on data mining," Informatics in Medicine Unlocked, volume 10, 2018, Pages 100-107.

[9] Ashwani Kumar, NeelamSahu, "Categorization of Liver Disease Using Classification Techniques," International Journal for Research in Applied Science & Engineering Technology, Volume 5 Issue V, May 2017.

[10] Kumar Y, Sahoo G, "Prediction of different types of liver diseases using rule based classification model," Technol Health Care. 2013;21(5):417-32. doi: 10.3233/THC-130742.

[11] A.Saranya1, G.Seenuvasan, "A Comparative Study of Diagnosing Liver Disorder Disease Using Classification Algorithm," International Journal of Computer Science and Mobile Computing, Vol. 6, Issue. 8, August 2017, pg.49 – 54.

[12] J. Kaur, "A comprehensive review on metabolic syndrome," Cardiology Research and Practice, vol. 2014, Article ID 943162, 21 pages, 2014.

[13] R. H. Eckel, S. M. Grundy, and P. Z. Zimmet, "The metabolic syndrome," The Lancet, vol. 365, no. 9468, pp. 1415–1428, 2005.

[14] Mercedes R. Carnethon**,** Catherine M. Loria**,** James O. Hill, Stephen Sidney, Peter J. Savageand Kiang Liu, "Risk Factors for the Metabolic Syndrome,"Diabetes Care 2004 Nov; 27(11): 2707-715. https://doi.org/10.2337/diacare.27.11.2.

[15] ApilakWorachartcheewan, Watshara Shoombuatong, Phannee Pidetcha, Wuttichai Nopnithipat, Virapong Prachayasittikul, and Chanin Nantasenamat, "Predicting Metabolic Syndrome Using the Random Forest Method," The Scientific World Journal, vol. 2015, Article ID 581501, 10 pages, 2015. doi:10.1155/2015/581501.

[16] Kathleen Davis, "Everything you need to know about hyperlipidemia," https://www.medicalnewstoday.com/articles.

[17] Blake Morris, "Lipid Profile (Triglycerides)" https://emedicine.medscape.com/article/2074115-overview.

[18] Verma A, Kaur I, Arora N, "Comparative analysis of information extraction techniques for data mining", Indian Journal of Science and Technology. 2016 Mar; 9(11). doi: 10.17485/ijst/2016/v9i11/80464.

[19] P. Lakhmi Prasanna, D. Rajeswara Rao, Y. Meghana, K. Maithri, T. Dhinesh "Analysis of supervised classification techniques", International Journal of Engineering & Technology, [S.l.], v. 7, n. 1.1, p. 283-285, Dec. 2017. ISSN 2227-524X. doi:http://dx.doi.org/10.14419/ijet.v7i1.1.9486.

[20] Mohamad Mumtazimah, Wan Nor Shuhadah Wan Nik, Zahrahtul Amani Zakaria, Arifah Che Alhadi, "An Analysis of Large Data Classification using Ensemble Neural Network", International Journal of Engineering & Technology, [S.l.], v. 7, n. 2.14, p.53-56,April2018. ISSN: 2227-524X. Available at: <https://www.sciencepubco.com/index.php/ijet/article/view/11155>. doi: http://dx.doi.org/10.14419/ijet.v7i2.14.11155.

[21] Buczak A L, Baugher B, Moniz L J, Bagley T, Babin S M, Guven E, "Ensemble method for dengue prediction", PLoS One. 2018Jan3;13(1):e0189988;doi:10.1371/journal.pone.01898 8.

[22] Kathleen H. Miao, Julia H. Miao and George J. Miao, "Diagnosing Coronary Heart Disease Using Ensemble Machine Learning", International Journal of Advanced Computer Science and Applications, Vol. 7, No. 10, 2016.

[23] Rosaida Rosly, Mokhairi Makhtar, Mohd Khalid Awang, Mohd Isa Awang, Mohd Nordin Abdul Rahman, Hairulnizam Mahdin, "Comprehensive study on ensemble classification for medical applications", International Journal of Engineering & Technology, 7 (2.14) (2018) 186-190.

[24] S., Nanda; M., Sukumar. "Detection and classification of thyroid nodule using Shearlet coefficients and support vector machine", International Journal of Engineering & Technology, [S.l.], v. 6, n. 3, p. 50-53, june 2017. ISSN 2227-524X. doi:http://dx.doi.org/10.14419/ijet.v6i3.7705.

[25] Nasser H.Sweilam, A.A.Tharwat, N.K.Abdel Moniem, "Support vector machine for diagnosis cancer disease: A comparative study", Egyptian Informatics Journal Volume 11, Issue 2, December 2010, Pages 81-92.

[26] H. I. Elshazly A. M. Elkorany A. E. Hassanien "Lymph diseases diagnosis approach based on support vector machines with different kernel functions" Computer Engineering & Systems 2014.

[27] Rupali R.Patil, "Heart Disease Prediction System using Naive Bayes and Jelinek-mercer smoothing", International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 5, May 2014.

[28] Aieman Quadir Siddique, Md. Saddam Hossain, "Predicting Heart-disease from Medical Data by Applying Naïve Bayes and Apriori Algorithm", International Journal of Scientific & Engineering Research, Volume 4, Issue 10, October-2013 ISSN 2229-5518.

[29] Ms. Ankita R. Borkar, Prashant R. Deshmukh, "Naïve Bayes Classifier for Prediction of Swine Flu Disease", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 5, Issue 4, April 2015 ISSN: 2277 128X.

[30] Saloni Aneja, Sangeeta Lal, "Effective Asthma Disease Prediction Using Naive Bayes - Neural Network fusion technique", 2014 International Conference on Parallel, Distributed and Grid Computing.