A Hybrid Approach to Classification of Polycystic Ovarian Syndrome Dataset

B.S. Anuhya[1], Manaswini Chilla[1], Dola Sai Siva Bhaskar Thota[1], Vikas B.[1]
[1] Department of Computer Science and Engineering
GITAM Institute of Technology, GITAM
Visakhapatnam, India.
(E-mail: anuhya2304@gmail.com)

Abstract— Data mining is an efficient technology which has the capability to convert raw, disconnected data into a meaningful format, also called as knowledge. The most advantageous feature in data mining is that its uses are not restricted to only one particular field. This evolutionary technology can be implemented to gain knowledge from the raw data acquired from various field of sciences and in this paper, a small part of bioinformatics is dealt upon. PCOS, acronym for Polycystic Ovarian Syndrome, is a very common yet neglected endocrinal disease prevalent among women of the 18-40 age group. In this paper, a hybrid approach has been taken up for affluent classification of the retrieved PCOS data set. The hybrid approach involves two conventional classification techniques, namely, Support Vector Machine (SVM) and Genetic Algorithm (GA), which have been merged together to improve the accuracy of the results obtained previously. Furthermore, adaptive boosting (AdaBoost), a boosting technique, has also been contrived to further increase the accuracy and precision percentages. The present research paper, on the whole, focuses on increasing the rightness during classification of PCOS data set through a new, advanced and hybrid approach.

Keywords— *Data Mining, Polycystic Ovarian Syndrome, Support Vector Machines, Cross Validation, Adaboost, Accuracy.*

I. INTRODUCTION

Polycystic Ovarian Syndrome is a chronic hormonal disorder which causes ovaries to enlarge in child bearing women [1]. According to statistics, there are one million cases of PCOS per year in India. PCOS can last for a few years or could be lifelong. Treatment for this chronic condition is available however, the condition cannot be cured.

Data mining is used to infer knowledge from huge datasets. This means that useful information can be extracted from datasets and converted to understandable structures [2]. It also helps in discovering recurring patterns and unknown interesting patterns in datasets. Major components of Data Mining include Classification, Association Rules and Sequence Analysis. It has six common classes of tasks: Anomaly detection, Association Rule Learning, Clustering, Classification, Regression and Summarization.

Classification is a classic data mining technique that is sometimes also based on Machine Learning. In simple words, it classifies each item in a dataset into a predefined set of classes or groups [3]. It does so by making use of mathematical techniques few of which are decision tree, neural networks, statistics and linear programming. Classification also helps to understand how to classify an item of a dataset. The most popular Classification Techniques are Decision Tree Algorithms, Support Vector Machines, K-Nearest Neighbor and Naive Bayes Classifiers.

Support Vector Machine is a powerful classification and regression algorithm based on statistical learning [4]. It performs classification by finding the hyperplane that maximizes the margin between two classes. The aim of SVM is to solve the problem of interest without solving a more difficult problem as an intermediate step.

Genetic algorithms were developed by John Holland and his students and colleagues at the University of Michigan, are search based algorithms based on the concepts of natural selection and genetics. Useful when the search space is very large and there are large number of parameters are involved [4]**.** They are randomized in nature and perform a lot better than random local search, as they use historical data.

The above two described common classification methods have been employed to come up with a hybrid approach as an enhancement to the previous research to further elevate the accuracy of predicting whether a patient is suffering from PCOS. In addition, AdaBoost has also been applied in order to amplify the accuracy percentages.

II. DATASET

The dataset for PCOS is a real-time data set that taken from a survey conducted among 119 women between the ages of 18 and 22. The dataset is primarily based on their lifestyle and food intake habits. The symptoms i.e. attributes are classified based on classification algorithms to predict whether the patient may have PCOS or not. The database consists of 119 samples with 18 attributes belonging to two different classes (maybe or maybe not). There are 14 binary attributes and 4 categorical attributes as shown below:

TABLE I. PCOS DATASET

S.NO.	ATTRIBUTE	VALUE
1	CLASS LABEL	MAYBE, MAYBE NOT (mb, mb n)
2	REGULARITY OF MENSTRUAL PERIODS	Yes (y), Infrequent menses (im), Irregular bleeding (ib), Heavy bleeding (hb)
3	WEIGHT GAIN	Yes(y), No (n)
4	EXCESS FACIAL OR BODY HAIR.	Yes (y), No (n)
5	DARK AREAS ON SKIN	Yes (y), No (n)
6	PIMPLES	Yes(y), No (n)
7	DEPRESSION AND ANXIETY	Yes(y), No (n)
8	HISTORY OF DIABATES AND HYPER TENSION	Yes(y), No (n)
9	BODY WEIGHT MAINTAINENCE	Yes(y), No (n)
10	OILY SKIN	Yes(y), No (n)
11	LOSS OF HAIR	Yes(y), No (n)
12	FREQUENT EATING PLACES	Hostel mess(hm), Campus canteen (cc)
13	REGULAR EXERCISE	Yes(y), No (n)
14	MENTAL STRESS DUE TO NEW ADMISSION IN HOSTEL	Yes(y), No (n)
15	MENTAL STRESS DUE TO PERSONAL PROBLEMS	Yes(y), No (n)
16	MENTAL STRESS DUE TO PEER PRESSURE	Yes(y), No (n)
17	MENTAL STRESS DUE TO CHANGE IN DIETARY HABITS	Yes(y), No (n)
18	FAST FOOD INTAKE	Every day(ed), Once in a week(w), Once in a month(m), Once in a year (y)

III. SUPPORT VECTOR MACHINES

SVM is one of a kind technique, a supervised learning algorithm, for data classification based on statistic study theory. The goal of the SVM algorithm is to determine a hyper plane that optimally separates two classes. There might be various hyper planes separating two classes.

However only a single hyper plane exists, that provides maximum margin between the two classes. An optimum hyper plane is determined using train data sets and its ability to classify is verified using test data sets. This particular algorithm has higher training speed when compared to other classification algorithms, such as neural network. SVM proves to be one of the most ideal algorithm when it comes to its excellent capability of generalization. [5]

In this algorithm, we plot each data item as a point in n-dimensional space in which n is number of features present, with the value of each feature being the value of a particular coordinate. After which, we perform classification by finding the hyper-plane that differentiate the two classes appropriately. In SVM, classification hypersurface is pointed out because of the induction of the support vectors, so the hypersurface basically reflects the relation between categorical attribute and condition attributes obtained by SVM which in turn promotes the classification [5]. Support Vectors are nothing but the co-ordinates of the individual observation and are the points that determine the limit on the width of the margin. Support Vector Machine is a line that best segregates the two classes.

The algorithm of the binary-class support vector machines is defined as follows [6]:
Let there are N training samples,

$$D = \{(a_i,b_i)^N\, i=1, A_j \in R^d,\ b_i \in \{+1,-1\} \qquad (1)$$

where, a_i denotes the input pattern for the i'th training sample, d denotes dimension of the input pattern, and b_i is the class of the i'th sample. We assume that the patterns having positive class and the patterns having negative class are linearly separable. The equation of the decision surface in the form of hyper plane that does the separation is given as:

$$(z^T, a_i) + c = 0 \qquad (2)$$

where, z is nothing but an adjustable weight vector and c is a bias.The constraint of the separating hyper plane can be written as:

$$(z^T, a_i) + c \geq 1 \quad i=1,2\ldots, N;\ b_i=+1 \qquad (3)$$

And

$$(z^T, a_i) + c \leq 1 \quad i=1,2\ldots, N;\ b_i=-1 \qquad (4)$$

Combining (3) and (4) the following equation is obtained.

$$b_i((z^T, a_i) + c) \geq 1;\ i=1,2\ldots,N; \qquad (5)$$

The Equation (1) defines the separation between the hyperplane for a particular weight vector (z) and bias (c). The SVM ultimately finds out the hyper plane with maximum margin of separation. In this situation, the decision surface is cited as optimal separating hyper plane.
The particular data point (a_i, b_i) satisfying (5) with equality sign is called support vector. The distance of a support vector from the optimal separating hyper plane is $\frac{1}{||w||}$. Hence, the optimum value of the margin of separation between two classes will be $\frac{2}{||w||}$. The optimal separating hyper plane minimizes to the function as follows:

$$r(z) = \frac{1}{2}z^T z. \qquad (6)$$

A. Kernel Functions

In support vector machine, a hyperplane or set of hyperplanes are constructed in a high- or infinite- dimensional space, which can be used for classification, regression, or other tasks [7]. Naturally, the best separation is obtained by the hyperplane that has the largest distance to the nearest training data points of any class, because larger is the margin the lower will be the generalization error of the classifier. While the actual problem may be defined in a finite dimensional space, it frequently happens that the sets to discriminate are usually not linearly separable in that space. Because of which, it was put forward that the original finite-dimensional space be mapped into a much larger-dimensional space, most likely making the separation easier in that space. To keep the computational load reasonable, the mapping used by the SVM schemes are designed to make sure that dot products can be calculated with ease in terms of the variables in the original space, by defining them in terms of a kernel function *K(x,y)* selected according to the problem [8]. The hyperplanes in the higher dimensional space are defined as the set of points whose inner product with a vector in that space is constant.

The kernel used for our hybrid approach towards predicting whether a patient has PCOS is the dot kernel.

Dot kernel: The dot kernel is nothing but the inner product of a and b. It is represented as

$$k(a,b)=a*b$$

where a and b are set of data points in our data set.

IV. GENETIC ALGORITHMS

Genetic algorithms (GAs) are best for searching global optimal values in complicated search space which might be multi-modal, multi-objective, non-linear, discontinuous, and highly restrictive space, along with the fact that they work with raw objectives, when compared with conventional techniques.
The above figure shows the general flow chart of GA and the main components that contribute to the overall algorithm. The operation of the GA starts with the first step called initialization which involves determining an initial population either randomly or by the use of some heuristics. The second step is nothing but the fitness assignment where fitness function is used to evaluate the members of the population and then they are ranked based on the performances. Once all the members of the population have been evaluated, the lower rank chromosomes are omitted and the remaining populations are used for reproduction. This is one of the most common approaches used for GA.

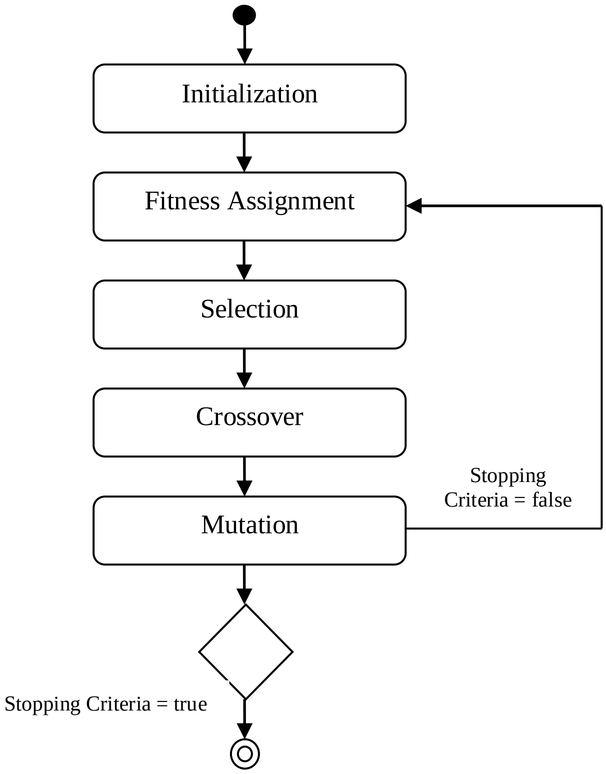


Fig. 1 General flowchart of Genetic Algorithm [9]

Another possible selection scheme is to use pseudo-random selection, allowing lower rank chromosomes to have a chance to be selected for reproduction. The crossover step randomly selects two members of the two members of the remaining population (the fittest chromosomes) and exchanges and mates them. The final step of GA is mutation. In this step, the mutation operator randomly mutates on a gene of a chromosome. Mutation is a crucial step in GA since it ensures that every region of the problem space can be reached [10]

V. BOOSTING

Boosting is a general method for improving the accuracy of any given learning algorithm. Boosting refers to a general and provably effective method of producing a very accurate prediction rule by combining rough and moderately inaccurate rules of thumb [11]. It is based on the observation that finding many rough rules of thumb can be a lot easier than finding a single, highly accurate classifier. [12]

VI. ADABOOST

The AdaBoost algorithm, introduced in 1995 by Freund and Schapire, solved many of the practical difficulties of the earlier boosting algorithms [11]. AdaBoost is a popular boosting technique which helps you combine multiple “weak classifiers” into a single “strong classifier”. A weak classifier is simply a classifier that performs poorly but performs better than random guessing. AdaBoost can be applied to any classification algorithm, so it’s really a technique that builds on top of other classifiers as opposed to being a classifier itself. There are two steps involved in Adaboost, one is training set selection where It helps you choose the training set for each new classifier that you train based on the results of the previous classifier. And the next step is Classifier Output Weights where It determines how much weight should be given to each classifier's proposed answer when combining the results. AdaBoost is fast, simple and easy to program. It has no parameters to tune (except for the number of round) [11].

VII. Performance Measures

The performance of the hybrid classification technique can be depicted by the following
Metrics:

A. Accuracy

The percentage of the test records that are properly classified by the classifiers is called accuracy [12].

Accuracy =

$$\frac{Occurrences\ of\ true\ positives + Occurrences\ of\ true\ negatives}{Occurrences\ of\ true\ positives + false\ \ negatives + true\ negatives}$$

B. Precision

It is the measure in which the fraction of true positives against all positive results is calculated [13].

Precision =

$$\frac{Occurrences\ \ of\ true\ positives}{Occurrences\ of\ true\ positives + Occurrences\ of\ \ false\ positives}$$

C. *Recall*

It is the ratio of the number of relevant tuples obtained to the total number of relevant tuples in the data set. It is usually expressed as a percentage.

$$\text{Recall} = \frac{X}{X+Y}(100)$$

X= number of relevant tuples obtained

Y=number of relevant tuples not obtained [14]

VIII. Proposed Algorithm

It has been seen in various papers that applying conventional classification techniques on required dataset, though is productive, but lacks in the degree of accuracy. Hence, in this proposed hybrid GA-SVM algorithm, an attempt has been made to improve the accuracy in classifying the Polycystic Ovarian Syndrome (PCOS) dataset when compared with applying only Genetic Algorithm or Support Vector Machine algorithm separately on the procured dataset.

A detailed explanation along with the step-wise algorithm and flow chart has been has been vividly specified here.

The retrieved PCOS Dataset is inputted and the class label is set appropriately. Next, the Genetic Algorithm is applied to a certain population chosen randomly from the dataset. Support Vector Machine is now applied to the linearly separable outputs to classify the tuples in the set. To determine the accuracy of this hybrid technique, cross validation is performed. To further enhance the accuracy of the algorithm, AdaBoost, a boosting technique, is applied while SVM is being implemented.

A. Flowchart

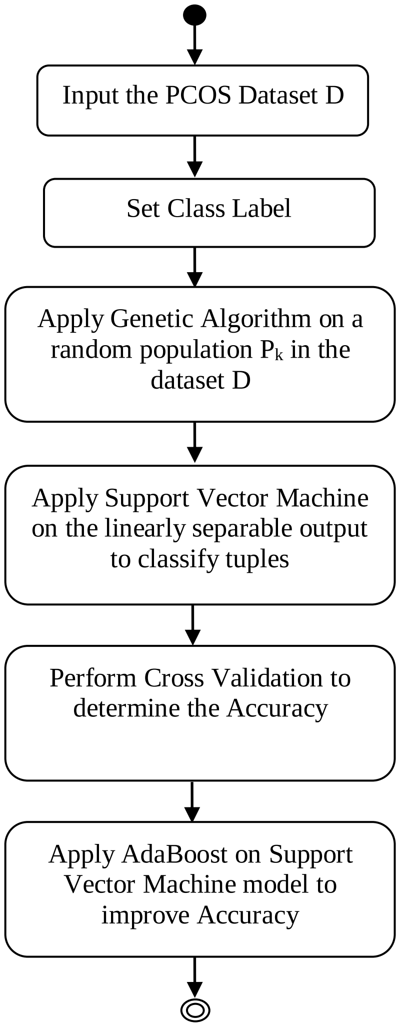

Input the PCOS Dataset D
Set Class Label
Apply Genetic Algorithm on a random population P_k in the dataset D
Apply Support Vector Machine on the linearly separable output to classify tuples
Perform Cross Validation to determine the Accuracy
Apply AdaBoost on Support Vector Machine model to improve Accuracy

Fig. 2 Proposed Flowchart

B. Algorithm

Input:

Let D be the PCOS Dataset containing m attributes and n instances/ number of individuals in the population.

x is the fraction of the population to be replaced by crossover in each iteration.

μ is the mutation rate.

λ is a tuning parameter for subset size.

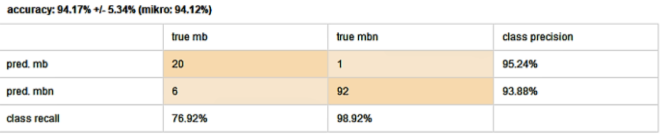
Algorithm:

```
GA(n, x, μ)
{
k := 0; // Initialise generation 0
$P_k$= a population of n randomly-generated individuals;
Compute fitness(i) for each i ∈ $P_k$;
do
{
    Select (1 − x) × n members of $P_k$ and insert into $P_{k+1}$; // Create generation k + 1
    Select x × n members of $P_k$; pair them up; produce offspring; //Crossover
    insert the offspring into $P_{k+1}$;
    Select μ × n members of $P_{k+1}$; //Mutate
    invert a randomly-selected bit in each;
    Compute fitness(i) for each i ∈ $P_k$; // Evaluate $P_{k+1}$
    k := k + 1;
}
while fitness of fittest individual in $P_k$ is not high enough;
return the fittest individual from $P_k$;
}

SVM(fittest individual from $P_k$)
{
    candidate = { closest pair from opposite classes }
    while there are violating points do
    //Find a violator
    candidate = candidate ∪ violator
    if any coefficient of a point p < 0 due to addition of the support vector c to Pk then
    candidate = candidate \ p
     repeat till all such points are pruned
    end if
    end while
}
Return classified data CD

Cross Validation( CD)
    {
    Divide CD into K roughly equal parts: training and testing data
    for each k=1,2…K, fit the model with a parameter λ to the other K-1 parts
    Compute the Error
    E(λ)=$\frac{1}{K}\sum_{k=1}^{K} E_k(\lambda)$
Do this for many values of λ and choose that which gives the smallest error
}

Adaboost(CD)
{
initialize the weight of each tuple in CD to 1/d;
for i=1 to k do //for each round;
            sample CD with replacement according to the tuple weights to obtain $CD_i$,
            use training set $CD_i$ to derive a model, $M_i$;
            compute error ($M_i$) the error rate of $M_i$
            if error(Mi) > 0.5 then
                        go back to step 3 and try again;
            endif
            for each tuple in $CD_i$ that was correctly classified do
                        multiply the weight of the tuple by error(Mi)(1-error(Mi)); //update weights
            normalize the weight of each tuple;
 endfor
initialise weight of each class to 0; //To use the ensemble to classify tuple, X
 for i=1 to k do // for each classifier:
            wi = $\log\frac{1-error(M_i))}{error(M_i)}$; // weight of the classifier's vote
            c = Mi(X); // get class prediction for X from Mi
            add wi to weight for class c
 end for
 return the class with the largest weight;
}
```

IX. Results

The figure 3 shows the confusion matrix for accuracy of the proposed hybrid approach to classify the PCOS data.

The accuracy comes out to be 94.17% which is more than accuracy obtained in correctly predicting the class label in the previous experiments.

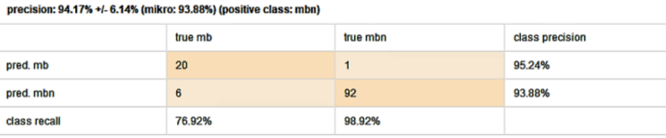


accuracy: 94.17% +/- 5.34% (mikro: 94.12%)

	true mb	true mbn	class precision
pred. mb	20	1	95.24%
pred. mbn	6	92	93.88%
class recall	76.92%	98.92%	

Fig. 3 Confusion matrix for accuracy

Figure 4 depicts the confusion matrix describing the precision of the hybrid algorithm which turns out to be 94.17%.

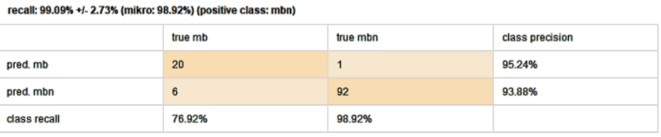


precision: 94.17% +/- 6.14% (mikro: 93.88%) (positive class: mbn)

	true mb	true mbn	class precision
pred. mb	20	1	95.24%
pred. mbn	6	92	93.88%
class recall	76.92%	98.92%	

Fig. 4 Confusion matrix for precision

Figure 5 depicts the confusion matrix describing the recall of the hybrid algorithm which turns out to be 99.09%.



recall: 99.09% +/- 2.73% (mikro: 98.92%) (positive class: mbn)

	true mb	true mbn	class precision
pred. mb	20	1	95.24%
pred. mbn	6	92	93.88%
class recall	76.92%	98.92%	

Fig. 5 Confusion matrix for recall

The area under curve (AUC) graph has also been shown in the figure 6. The graph is a kind of metric or representation which depicts the performance of a binary classification.

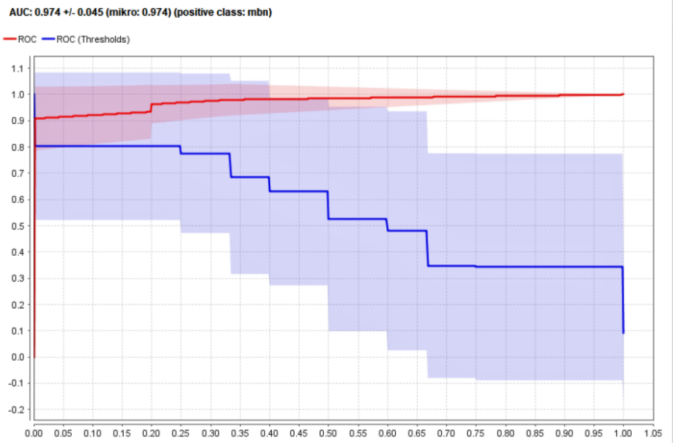


Fig. 6 AUC graph

X. Conclusion

In this paper, it is seen that an attempt has been made by the authors to increase the accuracy and hence the efficiency of the conventional classification techniques by following the mentioned hybrid approach. Through prior studies, it has come to notice that SVM and GA, though effective when applied individually, don't provide with optimal results. Hence, an effort has been made to merge these algorithms into one and as the result analysis above shows, the percentage of accuracy has drastically improved. The data set used in this research is a real time data acquired by setting up a survey in the near by localities. The classification of PCOS data can be made more accurate by the inclusion of clinical data and medical reports into this hybrid classification technique. This would, on the larger scale, help women to know more about PCOS, generate public awareness about the syndrome and act upon it at the earliest based on the prediction done.

Acknowledgment

We would like to pay our heartfelt gratitude to Dr. Vijaya Lakshmi Chandrasekhar, Department of Obst. & Gyn., GIMSR, GITAM (Deemed to be University), Visakhapatnam for her generosity in accumulating the data and apprising us about the relationships among various symptoms. We would like to take this opportunity to thank GITAM (Deemed to be University) Girls Hostel Kokila Sadan and the Chief warden Prof. T. Sita Mahalakshmi for allowing us to conduct the survey and permitting us to gather data for the survey.

References

[1] Vikas B, Sipra Sarangi, Manaswini Chilla, K Santosh Bhargav, B S Anuhya. (2017). A Literature Review on The Rising Phenomenon PCOS. *International Journal of Advances in Engineering & Technology,2(10), 216-224.*

[2] Vikas B, B.S.Anuhya, K Santosh Bhargav, Sipra Sarangi, Manaswini Chilla. (2017, June). Application of the Apriori Algorithm for Prediction of Polycystic Ovarian Syndrome (PCOS). *4th International Conference on Information Systems Design and Intelligent Applications, 2017.*

[3] Deeba, K., & Amutha, B. (2016). Classification Algorithms of Data Mining. *Indian Journal Of Science And Technology*, *9*(39).

[4] Cristianini, N., & Scholkopf, B. (2000). An Introduction To Support Vector Machines And Other Kernel-Based Learning Methods , Cambridge University Press, Cambridge, 2000, xiii+189 pp., ISBN 0-521-78019-5. *Robotica*, *18*(6), 687-689.

[5] Verma, G., & Verma, V. (2012). Role and Applications of Genetic Algorithm in Data Mining. *International Journal Of Computer Applications*, *48*(17), 5-8. http://dx.doi.org/10.5120/7438-0267

[6] Yan-Feng Fan, De-Xian Zhang, Hua-Can He, A New Classification Algorithm Research. *IEEE Xplore*, 07 January 2008, Beijing, China.

[7] Aniruddha Dey, Shiladitya Chowdhury, Manas Ghosh. Face Recognition using Ensemble Support Vector Machine. *IEEE Xplore*, 25 December 2017, Kolkata, India.

[8] Wu, C., Tzeng, G., & Lin, R. (2009). A Novel hybrid genetic algorithm for kernel function and parameter optimization in support vector regression. *Expert Systems With Applications*, *36*(3), 4725-4735.

[9] Jain, S. (2018). Introduction to Genetic Algorithm & their application in data science. Analytics Vidhya. Retrieved 17 March 2018, from https://www.analyticsvidhya.com/blog/2017/07/introduction-to-genetic-algorithm/

[10] Ab Wahab, M., Nefti-Meziani, S., & Atyabi, A. (2015). A Comprehensive Review of Swarm Optimization Algorithms. *PLOS ONE*, *10*(5), e0122827. http://dx.doi.org/10.1371/journal.pone.0122827

[11] Yoav Freund Robert E. Schapire. A Short Introduction to Boosting. *Journal of Japanese Society for Artificial Intelligence*, 14(5):771-780, September, 1999, Florham Park, NJ 07932 US.

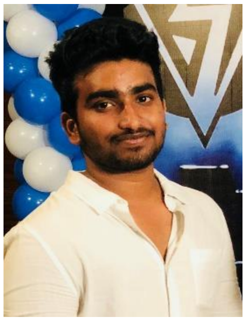
[12] Vikas B, Yaswanth D.V.S., Vinay W., Sridhar Reddy, Saranyu A.V.H. (2017, June). Classification of Hepatitis C Virus Using Case-Based Reasoning (CBR) with Correlation Lift Metric. *4th International Conference on Information Systems Design and Intelligent Applications, 2017.*

[13] Vladimirovich Shcherbakov, Adriaan Brebels. *A Survey of Forecast Error Measures Maxim, World Applied Sciences Journal 24 (Information Technologies in Modern Industry, Education & Society)*: 171-176, 2013 ISSN 1818-4952, Sep 25, 2013, Volgograd, Russia.

[14] Rapidminer Documentation retreived from https://docs.rapidminer.com/latest/studio/operators/modeling/predictive/support_vector_machines/support_vector_machine/

B S Anuhya is currently pursuing her Bachelor degree with the Department of Computer Science Engineering from GITAM (Deemed to be University), Visakhapatnam. Her research interests include data science, machine learning and cyber security.

Manaswini Chilla is currently pursuing her Bachelor degree with the Department of Computer Science Engineering from GITAM (Deemed to be University), Visakhapatnam. Her research interests include big data, network security and data mining.



Dola Sai Siva Bhaskar Thota is currently pursuing his Bachelor degree with the Department of Computer Science Engineering from GITAM (Deemed to be University), Visakhapatnam. His research interests include Deep Learning, Data Sciences, Machine Learning and Bioinformatics

Vikas B received the Bachelor in IT degree from the JNTUH, Hyderabad, in 2010 and the Master in Bioinformatics degree from the JNTUH, Hyderabad, in 2012. He is currently pursuing the Ph.D. degree with the Department of Computer Science and Engineering, GITAM (Deemed to be University), Visakhapatnam. His research interests include Deep Learning, Datamining, Bioinformatics, Information Security, and Data Sciences.