Feature Selection to Reduce Dimensionality of Heart Disease Dataset Without Compromising Accuracy

[1]Shiwani Gupta, [2]R. R. Sedamkar
[1]Assistant Professor, [2]Professor, Computer Engineering, Mumbai University
Thakur College of Engineering and Technology, Mumbai, India

Abstract *– Performance of machine classification is greatly affected by the selection of features and in medical field, accumulating data is a costly aspect. Even there is an increasing overfitting risk when no. of observations is insufficient and need for significant computation time when no. of features is more. Hence, it would be better if machines could extract most informative features i.e. medically high-risk factors to reduce the cost overhead on patients. Feature Selection is essential for simpler, faster, more reliable and robust machine learning models. Since wrapper-based methods are computationally expensive and filter-based methods are quicker, the authors claim through experimentation that filter based feature selection methods followed by wrapper can considerably reduce the size of feature set as well as enhance accuracy of prediction models onto high dimensional datasets without having to increase the number of instances. Results have been demonstrated on Arrythmia dataset from UCI Machine Learning Repository with 280 features and Z-Alizahdehsani dataset with 55 features.*

Keywords – *feature selection, heart disease, accuracy, filter, wrapper.*

I. INTRODUCTION

A. Preprocessing

Often raw data may be incomplete, inconsistent and may contain errors. To improve the quality of data and reduce its size, preprocessing is required. Data cleaning removes noise [1] and corrects inconsistency in data. It involves filling missing values as well. Further data transformation or normalization may be required to remove redundancy. Data reduction involves eliminating redundant features [1] through Principal Component Analysis, Linear Discriminant Analysis or Independent Component Analysis. Features with high percentage of missing values may be dropped. Highly correlated features can be identified and delineated. Features with zero/ low importance in tree-based model have no significance. Similarly Features with low unique value need to be dropped due to lack of information. Scaling is required to avoid attributes in greater numeric ranges dominate those in smaller numeric ranges [2]. Continuous attribute is converted to ordinal attribute to fit to Decision Tree (DT) algorithm [3]. Thus, features are to be standardized before feature selection [4].

B. Feature Selection

Dimensionality reduction leads to information loss whereas feature selection techniques are generally used for simplification of models to make them easier to interpret by researchers/users, take shorter training times, to avoid curse of dimensionality and provide enhanced generalization by reducing overfitting and hence variance. The central premise when using a feature selection technique is that data contains some features that are either redundant or irrelevant [5], and thus can be removed without much loss of information. Figure below explains that several subsets can be generated and evaluated to chose best according to its goodness.

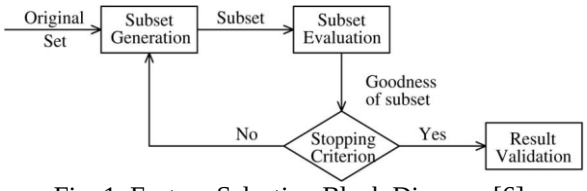


Fig. 1. Feature Selection Block Diagram [6]

These subsets are evaluated using distance measures, information measures, consistency measures and dependency measures.

Feature Selection algorithms are Filter based, Wrapper Based [7], Embedded and Hybrid. Filter based feature selection are further categorized as Basic, Multivariate and Statistical [8,9]. Wrapper methods are greedy, computationally intensive and

exhaustive. They look for subsets using Step Forward [10,4] or Step Backward [4] or Exhaustive search, build a model, and evaluate the subset for optimality through the score. This process is repeated until the performance of the model starts decreasing or till it keeps increasing or till pre determined number of features are extracted. Wrapper methods return the best performing feature set for that particular type of model. Embedded methods use a wrapper to consider interaction between feature and model but don't build a different model each time a different feature subset is picked. Embedded method is faster and cheaper than Wrapper and more accurate than Filter. The method is constrained to the limitation of associated algorithm. Types include Lasso Regularization and Decision Tree and Random Forest derived importance. Further Hybrid method can be employed to utilize advantages of both filter and wrapper methods. Other subset selection approaches include Simulated Annealing, Genetic Algorithm [2,11-14], Artificial Bee Colony metaheuristic [15] and Particle Swarm Optimization [12,16-17].

C. Applicability

Life threatening diseases particularly coronary artery disease [9,16,18], diabetes [13,18], Thyroid [10], liver [18], appendicitis [10] and cancer [1,16,18] have gained massive attention worldwide among individuals. The author has reviewed literature of more than 40 years to understand the developments in the field. The author inferences that cost could be saved, further leading to decision making if the large amounts of data are processed and analyzed. Poor clinical decisions due to lack of expertise can lead to loss of life which is unacceptable. Further, dimensionality reduction is essential to reduce the complexity of the algorithm with large number of features which further requires a greater number of instances for enhanced accuracy [25].

II. THEORETICAL FRAMEWORK

A. Filter

Filter methods rank features independent of others. There are various measures in Filter based methods as Correlation based, Consistency Based and Information Theory. They may select redundant features [1], hence data preprocessing is essential. The first step is to remove constant information which provides no/ minimal information since it has same value for all instances of the feature. This can be checked by checking the variance of feature values, if the variance is 0 then the feature values are redundant. There is possibility of feature values being quasi-constant i.e. the variance exceeds 0.8. Similarly post one hot encoding of large datasets or dataset with lots of categorical values, there is chance of duplicate rows. Hence transpose the dataset and perform same operations as for constant and quasi constant columns. Next step should be to identify correlated features since effective results are appreciated if features correlate highly to target and are uncorrelated to each other [8,19]. Pearson Correlation Coefficient and Mathews Correlation Coefficient [4] perform the task. Pearson Correlation is sensitive only to linear relationship. For Pearson correlation, drop feature with value close to 0. Further Statistical methods are fast but do not capture redundancy among features. These methods assign a score to each feature and thus rank them for inclusion or exclusion. Fisher score [4,16] and Chi square test [9] can be used to measure dependence among features. The methods are often univariate or consider the feature independently. Similarly, ANOVA parametric test identifies dependence among continuous variables. Information theory is used extensively as it can measure nonlinear relationship among features. Mutual Information (Information Gain) [3,17,20-23] is used a lot in literature to identify how much knowing one variable reduces the uncertainty of other. Mutual Information is inconvenient to compute for continuous variables.

TABLE I. FEATURE SELECTION METHODS

Methods	Methodology	Merits	Gaps	Solution to overcome Gap
Feature Selection (Filter Based – Statistical and Correlation based)	Pearson Correlation, Mutual Info, Kendall's correlation, Spearman correlation, Chi square, Fisher score	Can use for preprocessing	Select redundant variables	Preprocessing prior to Filter based Feature Selection
Feature Selection (Wrapper Based – Deterministic and Randomized)	Backward Elimination, Forward Selection	Better results since tuned to classifier	Overfitting with insufficient sample size, time complexity with more no. of features	Utilizing Filter based prior to Wrapper based to reduce time complexity

A. Wrapper

Wrapper methods consider the selection of a set of features as a search problem, where different combinations are prepared, evaluated and compared to other combinations. Different combinations can be tried using Relief [17,19], Gain Ratio and Entropy [1,21]. They detect the possible interactions between variables. The search process may be methodical such as a best-first search [8], it may stochastic such as a random hill-climbing algorithm, or it may use heuristics, like forward and backward passes to add and remove features. Step Forward method evaluates all subsets with single feature then selects best feature. Now subsets with the best feature combined with another feature are evaluated to find best subset of 2 features. We repeat this till performance keeps enhancing. The performance metric used for evaluation is ROC AUC for classification and RSquared for prediction. Step Backward method [1] works opposite. It begins with all the features, removes 1 in each subset and evaluates these subsets, repeats this process till performance starts degrading. The main control issue is to decide when to stop the algorithm. In Machine Learning, we use cross validation and in statistics some criteria are optimized.

B. Hybrid

Recursive Feature Elimination commonly used with Support Vector Machine [24] or Random Forest to repeatedly construct a model and remove features with low weights. It is a greedy optimization algorithm which aims to find the best performing feature subset. It repeatedly creates models and keeps aside the best or the worst performing feature at each iteration. It constructs the next model with the left features until all the features are exhausted. It then ranks the features based on the order of their elimination. Pruning the search space can be done through Forward selection procedure combined with linear least square estimation algorithm [24].

C. Embedded

LASSO regularization adds penalty on different parameters to reduce their freedom making the model to fit noise of training data and thus generalize well on test data. We can ascertain that if penalty is too high, important features are dropped since the performance of model drops. For regression, the coefficients of predictors are proportional to how much it contributes to target variable. These methods work under the assumption that there is linear relationship between explanatory and predictor variable, explanatory variables are independent, normally distributed [13] and scaled. Thus, the features whose coefficients are more than mean of all coefficients are selected. The variants include L1(Lasso), L2(Ridge) and L1/L2(Elastic Net). L1 might shrink some parameters to zero but in L2, parameters never shrink to zero but approach it. Improvements to Lasso include Bolasso which bootstraps samples. For Tree derived importance algorithms, a feature is more important if it reduces the impurity more, example being Regularized Random Forest.

III. PROPOSED WORK

The literature reviewed states that high dimensional datasets require more no. of instances. Collecting data is a costly aspect. Hence, it is better to identify informative features, and thus reduce complexity in order to save computation. Figure below shows the proposed architecture:

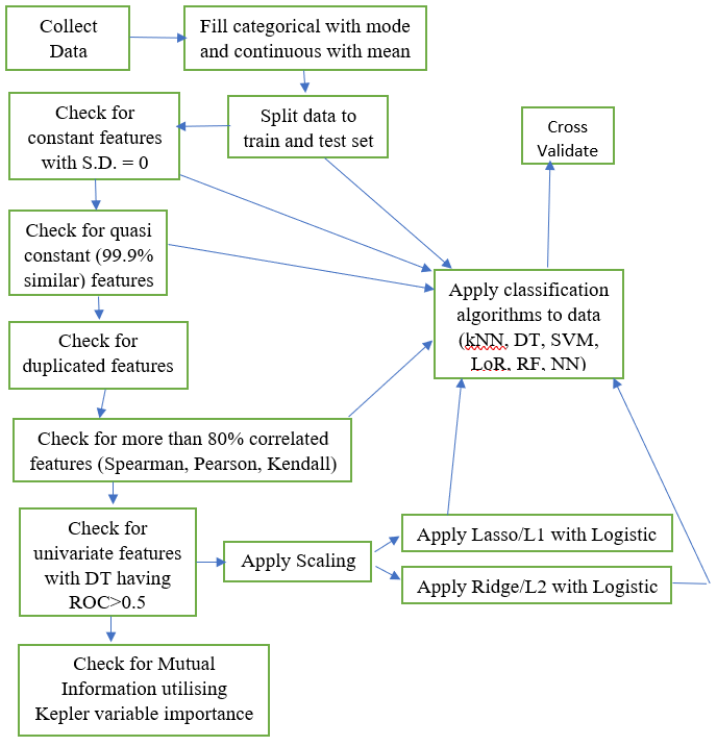


Fig. 2. Proposed Work Steps

IV. RESULTS AND FINDINGS

The authors have worked on high dimensional heart disease - Arrythmia dataset with 280 features. The data comprises of problems related to heart rhythm caused by improper working of electrical impulses. Since the data is high dimensional, directly developing classification on the data provided following results with NN and RF giving best accuracy.

Since the model is predicting Heart disease too many type II errors are not advisable. Results demonstrate high sensitivity with NB and kNN, which is of more importance to healthcare data than specificity. Hence all four cells in confusion matrix – True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN) are utilised for computing different performance measures.

Accuracy= (TP+TN)/(TP+FP+TN+FN) …..(1)
Sensitivity = Recall = TP / T = TP / (TP + FN) .....(2)
Specificity = TN / N = TN / (TN + FP) ….(3)
Precision = TP / P = TP / (TP + FP) ….(4)
In all NN have demonstrated better results.

TABLE 2 PERFORMANCE COMPARISON ON RAW DATA

Algorithm	Accuracy	Sensitivity/Recall	Specificity	Precision
Logistic Regression	61.76%	76.39%	60.94%	68.75%
Gaussian Naïve Bayes	61.76%	**95.83%**	46.88%	66.99%
Decision Tree	69.85%	76.39%	57.81%	67.07%
Random Forest	**72.05%**	75.0%	70.31%	73.97%
K Nearest Neighbor	63.23%	90.28%	34.38%	60.75%
Support Vector Machine	52.9%	69.44%	65.63%	69.44%
NN	**74.26%**	71.64%	**76.81%**	**75%**

Since the accuracy is low, without pre-processing the data; following feature selection algorithms were applied in sequential order to reduce dimensionality of data and enhance performance.

- Filter based
  - Basic (Constant, Quasi Constant, Duplicated)
  - Correlation (Pearson, Kendall, Spearman)
  - Statistical (Univariate, Mutual Information
- Embedded (Lasso)
- Wrapper based (Sequential Forward Search)

Six classification algorithms tested the accuracy on processed data. Results can be seen in Table below:

TABLE 3 ACCURACY COMPARISON WITH FEATURE SELECTION

Type	Step	LoR	KNN	SVM	DT	RF	NN
Filter Based (Basic)	Remove constant, quasi constant and duplicated features	61.03	63.23	63.47	69.12	75	75.74
Filter Based (Correlated)	Remove Correlated Features (Pearson)	68.38	63.23	66.17	69.12	70.59	74.26
	Remove Correlated Features (Kendall)	61.03	63.23	60.29	65.44	72.06	72.79
	Remove Correlated Features (Spearman)	68.38	65.44	68.38	63.23	75	69.85
Statistical	Univariate	68.38	63.23	68.38	65.44	70.59	72.06
Embedded	Apply Lasso	**69.85**	65.4	**72.05**	**75**	69.85	52.94
	Apply Ridge	63.23	63.23	60.29	63.23	72.06	
Wrapper	Sequential Feature Selection	66.43	**74.26**	65.44	69.11	**91.88**	45.48

Thus, we see that Wrapper based Sequential feature selection with Random Forest gives good accuracy enhancement. Normalising the data before feature selection further enhances accuracy of kNN to 68.4% post Ridge, Lasso to 72.8% with SVM and since DT doesn't require scaling, results don't improve. The plot below shows accuracy comparison for different classification models post filtering by different strategies.

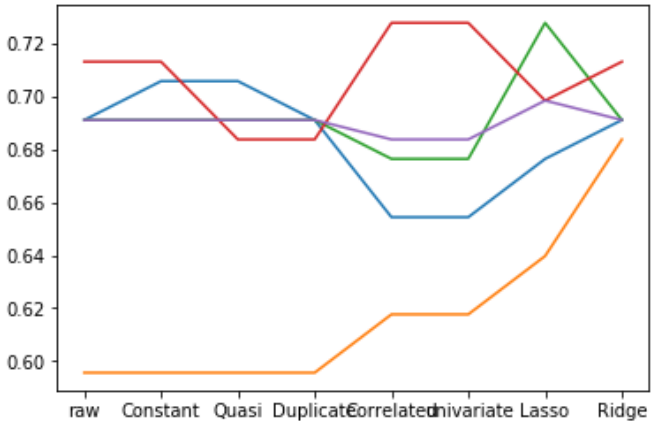


Fig. 3. Accuracy comparison for DT (blue), kNN (orange), SVM (green), RF (red), LoR(mauve)

Further we experimented applying dimensionality reduction techniques PCA and LDA with not much enhancement in accuracy.

With Exploratory data analysis, it was found that 1 feature has a lot of missing values, so was dropped. Further some features had constant values which were not contributing to the information. Correlated features bring redundant information; thus, they can be removed to reduce dimensionality. Lasso Regression performs both variable selection and regularisation to enhance accuracy and interpretability. Sequential Forward Feature selection is better than Sequential backward since only 7 features are selected out of 99 by SFS. Following table shows the no. of features selected by each algorithm without compromising accuracy.

TABLE 4 SIZE OF FEATURE SUBSET WITH FEATURE SELECTION

Step	**Features remaining**
Raw Data	280
Missing value col dropped	279
Constant col dropped	257
Quasi constant col dropped	214
Duplicated col dropped	214
Pearson Correlated col dropped	157
Kendall's Correlated col dropped	180
Spearman Correlated col dropped	147
Univariate columns dropped	147
Columns carrying mutual information dropped	78
Features removed through Lasso regularisation	99
Features selected through SFS (RF)	7

To enhance the accuracy further cross validation was applied on to NN result with increased accuracy of 75.29% mean with 6.5% variance from 72.79%. If you perform feature selection on all of the data and then cross-validate, then the test data in each fold of the cross-validation procedure was also used to choose the features and this is what biases the performance analysis. This means that feature

selection is to be performed on the prepared fold right before the model is trained.

Author has also applied feature selection on Z-AlizahdehSani dataset comprising of 55 features contributing to Coronary Artery disease (CAD) using:

Parametric tests as

Pearson Correlation = cov(x,y) / (sd(x) * sd(y))

….(5)

Fisher score Correlation = between class scatter/ within class scatter

Non parametric tests as

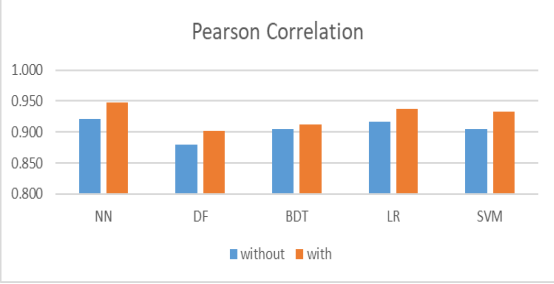
Kendall's tau Correlation = (# of concordant pairs - # of discordant pairs) /(n)(n-1)/2 ….(6)

Spearman's Rho Correlation = $\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$

….(7)

Each one of these reduced the dataset from 55 to 34 features with 21 having 0 correlation. Table below shows classification result comparison for complete and reduced dataset on several classification models. Positive value of Pearson Correlation Coefficient shows that features are linearly related. Fisher score is a derivative of Newton method. Kendall has smaller values than spearman with better statistical properties, though Spearman is widely used.

Further graph below shows that Pearson correlation and fisher score statistical test for feature selection show increased AUC with NN, DF, DT, RF and LoR, though mutual information and chi square did not give better results.



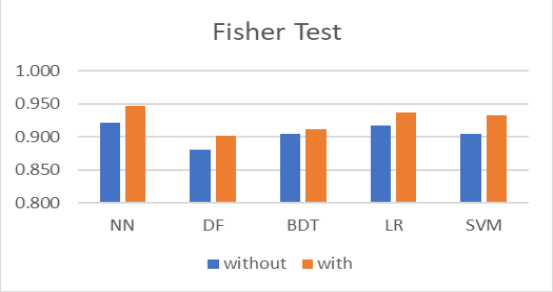


Fig. 4. Applying Feature selection enhances AUC

IV. CONCLUSION AND FUTURE SCOPE

The motivation behind this entire work was the available literature where different people worked upon different sized datasets with differing accuracy. Hence, I thought of picking two high dimensional datasets and coming up with reduced dimensions. Further, picking up a classification algorithm that would provide best performance for diagnosis. There is a lot of work in progress which can not be included in this paper due to lack of time. Further parameter tuning can enhance the performance and provide enhance generalisation capability. The work is aimed to reduce clinical costs by identifying informative features and thus is able to achieve so. Thus, it has been proven that Embedded methods have better accuracy than Filter.

ACKNOWLEDGMENT

I am grateful to many of my ex and current students, who have helped me understand and build experimentation by providing me courses and code when required.

REFERENCES

[1] H. M. Lee, C. M. Chen, J. M. Chen and y. L. Jou, "An efficient Fuzzy classifier with Feature selection based on Fuzzy entropy", *IEEE Transactions on Systems, Man and Cybernetics*, Vol 31, No. 3, Jun 2001.

[2] L. Huang, and C. J. Wang, "GA based feature selection and parameter optimization for support vector machine", *Elsevier Expert system with applications* 31 (2006) 231-240.

[3] J. Catlett, "On changing continuous attributes into ordered discrete attributes", SpringerLink, Jun 2005.

[4] S. M. Saqlain, M. Sher, F. A. Shah, I. Khan, M. U. Ashraf, M. Awais, A. Ghani, "Fisher score and Matthews correlation coefficient-based feature subset selection for heart disease diagnosis using support vector machines", *Springer Nature* 2018.

[5] Aha, "Tolerating noisy, irrelevant and novel attributes in Instance Based Learning Algorithms", Elsevier ScienceDirect, Feb 1992.

[6] H. Liu, and L. Yu, "Towards integrating feature selection algorithms for classification and clustering" *IEEE Transactions on knowledge and data engineering*, Vol. 17, No. 4, Apr 2005.
[7] R. Kohavi, G. H. John, "Wrappers for Feature Selection", Elsevier Artificial Intelligence (1997) 273-324.
A. Pandey, P. Pandey, K. L. Jaiswal, and A. K. Sen, "Datamining clustering techniques in prediction of heart disease using Attribute selection method", *International Journal of Science, Engineering and Technology Research* (IJSETR) Volume 2, Issue 10, October 2013.
[8] Peterkova, M. Nemeth, G. Michalconok, and A. Bohm "Computing Importance Value of Medical Data Parameters in Classification Tasks and Its Evaluation Using Machine Learning Methods", *Springer* 2019.
[9] M. L Raymer, W. L. Punch, E, D, Goodman, L. A. Kuhn, and A. K. Jain, "Dimensionality reduction using Genetic Algorithms", *IEEE Transaction on Evolutionary Computation*, Vol 4, No. 2, Jul 2000.
[10] S. Bhatia, P. Prakash, G. N. Pillai, "SVM based decision support system for heart disease classification with integer coded genetic algorithm to select critical features", WCECS, October 22 - 24, 2008, San Francisco, USA.
[11] Babaoglu, O. Findik, E. Ulker, "A comparison of feature selection models utilizing binary particle swarm optimization and genetic algorithm in determining coronary artery disease using support vector machine", *Elsevier Expert system with applications* 37 (2010) 3177-3183.
[12] T. Santhanam, M. S. Padmavathi, "Application of K-means and Genetic algorithm for dimension reduction by integrating SVM for diabetes diagnosis", *ScienceDirect Elsevier Procedia Computer Science* 47 (2015) 76-83.
[13] Yang, V. Honavar, "Feature subset selection using Genetic Algorithm", Kluwer Academic Publishers 1998.
[14] Subanya, Dr. R. R. Rajalaxmi, "Feature selection using Artificial Bee Colony for Cardiovascular Disease classification", 2014 *International Conference on Electronics and Communication System.*
[15] Y. Li, G. Wang, H. Chen, H. Dong, X. Zhu, S. Wang, "An improved Particle warm Optimization for Feature Selection", *ScienceDirect.com Journal of Bionic Engineering* 8 (2011) 191-200.
[16] Kononenko, "Estimating Attributes: Analysis and Extensions of Relief" CiteSeerX, 1994. Sequential
[17] M. Zhao, C. Fu, L. Ji, K. Tang, M. Zhou, "Feature selection and parameter optimization for support vector machines: A new approach based on genetic algorithm with feature chromosomes", *Elsevier Expert system with applications* 38 (2011) 5197-5204.
[18] M. A. Hall, "Correlation based Feature selection for Machine Learning", Univ. of Waikato, Apr 1999.
[19] Radha R., Murlidhar S., "Removal of redundant and irrelevant data from training datasets using speedy feature selection method", *International journal of Computer science and mobile computing* Vol 5, Issue 7, Jul 2016, pg. 359-364.
[20] T. T. Zhao, Y. B. Yuan, Y. J. Wang, J. Gao, P. He, "Heart disease classification based on feature fusion", IEEE International conference on Machine Learning and cybernetics, Ningbo, China 9-12 July 2017.
[21] P. Zhang, W. Gao, G. Liu, "Feature selection considering weighted relevancy", *Springer Nature* 2018.
[22] S. Jiang , K. S. Chin , G. Qu , K. L. Tsui , "An Integrated Machine Learning Framework for Hospital Readmission Prediction", *Knowledge-Based Systems* (2018).
[23] Z. Mao, "Feature subset selection for support vector machines through discriminative function pruning analysis", *IEEE Transactions on Systems, Man and Cybernetics*, Vol 34, No. 1, Feb 2004.
[24] S. Gupta, R. R. Sedamkar, "Apply Machine Learning for Healthcare to enhance performance and identify informative features", IEEE INDIACom; 6th International Conference on "Computing for Sustainable Global Development", 13th - 15th March, 2019, BVICAM, New Delhi (INDIA).
[25] https://udemy.com/feture-selction-for-machine-learning.

Ms. Gupta was born in India in 1981 has completed M.Tech in C.S.E. from Lucknow in 2009 and B. Tech in C.S.E. from Lucknow in 2003. She is currently pursuing Ph. D. in Tech. from Mumbai University. Her area of interest include Machine Learning, Artificial Intelligence, Biometrics, Algorithms etc.

She has over 15 years of teaching experience and is currently working as Deputy HOD in Computer Engineering Department in Mumbai. She has over 40 publications in reputed journals and conferences

Dr. R. R. Sedamkar was born in India in 1967 has completed Ph.D. in Technology from NMIMS Univ. in 2010 , M.E. in C.S.E. from Gulbarg Univ. in 2000 and B. E. in C.S.E. from Gulbarg Univ. in 1992. His areas of interest include Image Processing, Wireless Networks, Data Compression etc. He has over 23 years of teaching experience and is currently working as Professor and Dean Academic in TCET Mumbai. He has 8 research scholars working under him. He has over 40 publications in reputed journals and conferences.