ISSN No:-2456-2165

Feature Extraction and Selection Techniques for High-Dimensional Data

Chava Mojesh Chowdary
Dept of Information Technology
National Institute of Technology, Raipur, India

Roadmap:

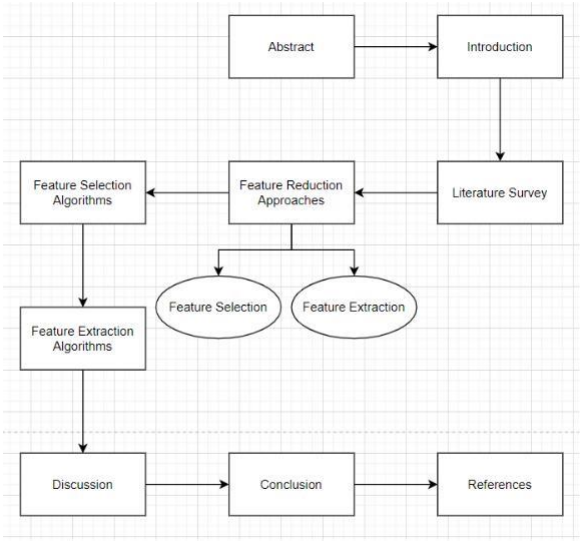

Fig. 1: Road map

Abstract:- As a preprocessing step, dimensionality reduction from high-dimensional data helps reduce unnecessary data, enhance learning accuracy, and improve result comprehensibility. However, the recent growth in data dimensionality offers a serious challenge to the efficiency and efficacy of many existing feature selection and feature extraction approaches. Dimensionality reduction is an essential topic in machine learning and pattern recognition, and numerous algorithms have been presented. In this research, certain commonly used feature selection and feature extraction approaches are examined to see how well they may be utilized to improve the performance of learning algorithms and, as a result, the predicted accuracy of clas N Ssifiers. A brief examination of dimensionality reduction approaches is offered to determine the strengths and limitations of various commonly used dimensionality reduction methods.

I. INTRODUCTION

In high-dimensional data analysis, visualization, and modeling, dimensionality reduction is a common preprocessing technique. Feature Selection is one of the easiest approaches to minimizing dimensionality; it picks only input dimensions with the necessary information for addressing the problem. Feature Extraction is a more broad strategy that involves attempting to build a transformation of the input space onto a low-dimensional subspace that maintains the majority of the important data. The goal is to increase performance, such as predicted accuracy, visualization, and comprehensibility of learned knowledge, using feature extraction and selection algorithms alone or in combination. Features might be classified as important, irrelevant, or redundant in general. A subset of accessible feature data is chosen for the learning algorithm during the feature selection procedure. The best subset is one with the fewest dimensions that contribute the most to learning accuracy.

The benefit of feature selection is that crucial information about a particular feature is not lost. Still, if just a limited number of features are needed, and the original features are quite diverse, there is a risk of information being lost since certain features must be excluded. Dimensionality reduction, also known as feature extraction, on the other hand, allows the size of the feature space to be reduced without losing information from the original feature space. One disadvantage of feature extraction is that the linear combination of the original characteristics is typically unintelligible, and information about how much each original feature contributes is frequently lost.

mRmR, CMIM, RELIEF, Correlation Coefficient, INTERACT, BW-ratio, GA, SVM-REF, PCA (Principal Component Analysis), Non-Linear Principal Component Analysis, Independent Component Analysis, and Correlation based feature selection are just a few of the techniques that have been developed. Given the large variety of available feature selection and feature extraction algorithms, it's important to have criteria to rely on when deciding which approach to apply in certain scenarios. A brief survey of these techniques is conducted based on a literature review to determine the suitability of various feature selection and feature extraction techniques in specific situations based on experiments conducted by researchers to determine how these techniques to aid in improving the predictive accuracy of classification algorithms. We introduce alternative dimensionality reduction strategies to interested readers in this paper.

II. LITERATURE SURVEY

In the medical profession, dimensionality reduction approaches have become a clear requirement (automated application). In today's world, a tremendous amount of data is created in the medical field. This covers a patient's symptoms and several medical test findings that may be generated. The terms "feature," "input variables," and "attributes" are interchangeable. The characteristics in medical diagnostic examples might include symptoms, which are factors that categorize a patient's health state (e.g., diabetic retinopathy symptoms of Dry or Wet Age-related macular degeneration (AMD)). This section

ISSN No:-2456-2165

presents a literature analysis of various commonly used feature selection and feature extraction approaches for ophthalmologists in the detection and diagnosis of numerous eye illnesses (glaucoma, diabetic retinopathy, and, notably, automated detection of age-related macular degeneration). The primary goal of this paper is to raise awareness among practitioners about the advantages and, in some circumstances, the necessity of using dimensionality reduction approaches. It is necessary to be aware of the many advantages of dimensionality reduction approaches to profit from them to enhance the accuracy of learning algorithms.

- It decreases the feature space's dimensionality to minimize storage needs and speed up the process.
- It eliminates data that is redundant, useless, or noisy.
- The immediate implications for data analysis activities include reducing the time it takes for learning algorithms to execute.
- Increasing data quality
- Improving the accuracy of the resulting model.
- Reduce the number of features in the feature set to save time and resources during the next data collection cycle or usage.
- To enhance prediction accuracy, performance must be improved.
- Data visualization or data comprehension to learn more about the process that created the data.

A review of relevant work on learning from noisy data was reviewed, suggesting that feature extraction is used as a preprocessing step to reduce the impact of class noise on the learning process. Filtering procedures specifically handle noise. Many filtering techniques have been summarized that researchers have found to be beneficial. On the other hand, the same researchers have identified several practical challenges with filtering algorithms. One difficulty is that without the assistance of an expert, distinguishing noise from exceptions (outliers) is difficult. Another issue is that a filtering algorithm may employ a predicted amount of noise as an input parameter, which is seldom known for specific datasets. Feature extraction approaches (using PCA) are preferable for noise-tolerant procedures since they reduce implicit overfitting inside learning algorithms. Using feature extraction techniques before supervised learning reduces the detrimental impact of the existence of mislabeled occurrences in the data.

A diabetes diagnosis strategy based on artificial neural networks (ANN) and a feature set derived from singular value decomposition (SVD) and Principal Component Analysis (PCA) have been suggested. The results of the experiments reveal that the ANN-SVD+PCA combination is a viable diabetes diagnosis method with low computing cost and good accuracy. Because of the noisy data, feature extraction approaches were far more suited for the automated identification of ophthalmology illnesses than feature selection methods because most biological datasets contain noisy data rather than useless or redundant data.

III. DIMENSIONALITY REDUCTION APPROACHES

Due to the high computational cost and memory utilization of high-dimensional data, classification algorithms struggle with it. Feature extraction (known as dimensionality reduction explicitly or feature transformation) and feature selection are two dimensionality reduction approaches (FS).

The benefit of FS is that no information regarding the relevance of a particular feature is lost. However, if a limited set of features is required and the original features are quite varied, information may be lost since certain characteristics must be excluded during the feature subset selection process. On the other hand, feature extraction allows the size of the feature space to be reduced without losing a lot of information from the original feature space. The decision between feature extraction and feature selection methods is determined by the application's unique data type domain.

A. Feature Selection:

High-dimensional data contains potentially irrelevant, deceptive, or redundant characteristics, resulting in a larger search area, making it more difficult to interpret data and thus not aiding the learning process. Selecting the best characteristics from all the features that may be used to distinguish classes is known as feature subset selection. The feature selection algorithm (FSA) is a computational model triggered by relevance criteria. In general, feature selection is referred to as a search problem that is based on a set of assessment criteria. The search structure of feature selection algorithms may be classified into three types: exponential, sequential, and random. (ii) Generation of successors (subset): To create successors, five distinct operators can be used: Forward, Backward, Weighted, Compound, and Random. (iii) Evaluation Measure: Probability of Error, Divergence, Interclass Distance Dependence, Consistency Evaluation, Information, or Uncertainty and may all be used to assess successors.

Filters, wrappers, and embedded/hybrid approaches are the three kinds of feature selection methods. Because the feature selection process is tuned for the classifier to be employed, wrapper approaches outperform filter methods.

On the other hand, Wrapper techniques are too expensive to utilize for vast feature spaces due to their high computational cost. Each feature set must be validated with the trained classifier, slowing down the feature selection process. Filter techniques have a lower computing cost and are speedier, but they have lower classification reliability than wrapper approaches, which are better suited to high-dimensional data sets. Hybrid/embedded solutions have recently been created, including the benefits of filters and wrapper approaches. A hybrid technique employs a feature subset's independent test and performance assessment function. Filter techniques are further divided into two categories: feature weighting algorithms and subset search algorithms. Feature weighting algorithms give each

feature weight and rank it according to its relevance to the goal notion.

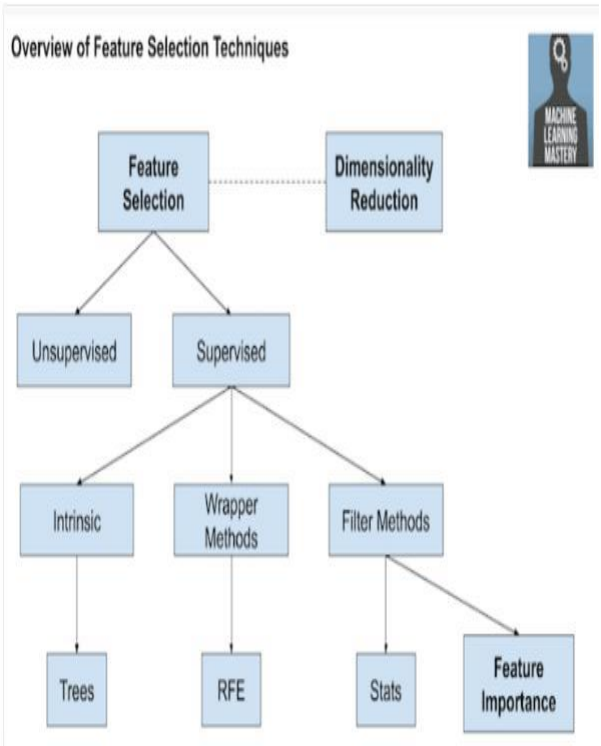

Fig. 2: Overview of Features Selection Techniques

- **Wrapper**: Looks for feature subsets that perform well.
  - RFE
- **Filter**: Choose subsets of features based on their relationship with the target.
  - Statistical Methods
  - Feature Importance Methods
- **Intrinsic**: Algorithms that make automatic feature selection during the training.
  - Decision Trees

B. Choosing feature Method:

The more information about a variable's data type, the easier it is to pick a statistical measure using a filter-based feature selection approach. The variables supplied as input to a model are known as input variables. These are the variables we want to shrink in size during feature selection. The output variables, also known as the response variables, are the ones a model is supposed to predict.

C. Numerical Input, Numerical Output:

This is a numerical input variable regression predictive modeling issue. A correlation coefficient, such as Pearson's, is used for a linear correlation, while rank-based approaches are used for a nonlinear correlation.

- Pearson's correlation coefficient (linear).
- Spearman's rank coefficient (nonlinear)

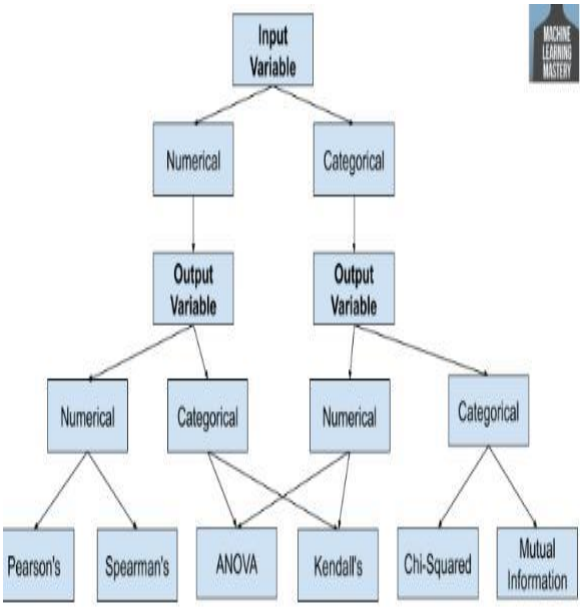

Fig. 3: Numerical Input, Numerical Output:

D. Numerical Input, Categorical Output:

This is a numerical input variable categorization predictive modeling issue. This is perhaps the most prevalent type of categorization issue. The most prevalent strategies are correlation-based once again, but this time they must account for the category goal.

- ANOVA correlation coefficient (linear).
- Kendall's rank coefficient (nonlinear).

E. Categorical Input, Numerical Output:

This is a categorical input variable regression predictive modeling issue. This is an unusual instance of a regression problem (e.g., you would not encounter it often). However, the same "Numerical Input, Categorical Output" approaches (explained above) may be used reversely.

F. Categorical Input, Categorical Output:

This is a categorical input variable categorization predictive modeling issue. The chi-squared test is the most popular correlation metric for categorical data. Mutual information (information gain) from the discipline of information theory can also be used.

- Chi-Squared test (contingency tables).
- Mutual Information.

G. Feature Selection Algorithms:

The Chi-squared test is the most often used statistical test for determining how much a feature's occurrence deviates from the predicted distribution if the feature's occurrence is assumed to be independent of the class values. Euclidean The root of square discrepancies between the coordinates of two objects are investigated by distance. The advantage of this approach is that adding additional items to the study, which may be outliers, does not affect the distance. However, differences in size among the dimensions from which the distance is derived can significantly impact Euclidean distance. The t-test determines if the two groups'

averages are statistically different from one another. This approach is recommended anytime two groups' averages must be compared, and it is particularly well suited for the posttest-only two-group randomized experimental design. Information Gain (IG) is a metric that compares the increase in entropy when a feature is present vs when it is not.

This is the application of more broad approaches, such as informational entropy measurement, to the challenge of determining the importance of a feature in feature space. Correlation-Based Feature Selection (CFS) looks for feature subsets based on how redundant the features are. The assessment aims to uncover subsets of characteristics that are substantially associated with the class separately but have minimal inter-correlation. The importance of a set of traits increases as the correlation between them and the class increases and reduces as the inter-correlation increases. CFS is commonly used in conjunction with search algorithms such as forward selection, backward elimination, bi-directional search, best-first search, and genetic search to discover the optimal feature subset.

The simplest greedy search technique is Sequential Forward Selection (SFS). When the ideal subset contains a minimal number of characteristics, SFS performs well. The fundamental drawback of SFS is that it is impossible to delete features that become obsolete when new ones are added. Sequential Backward Elimination (SBE) is the polar opposite of Sequential Forward Elimination (SFS). When the feature subset includes many features, SBE performs well. SBE's fundamental flaw is its inability to reconsider the use of a feature after it has been removed. LRS (Plus-L Minus-R Selection) is a hybrid of SFS and SBE. With certain backtracking capabilities, it aims to compensate for the shortcomings of SFS and SBE. Its fundamental flaw is its lack of theory to assist in forecasting the best values of L, and R. Individual feature selection techniques have the drawback of only capturing the relevance of characteristics to the goal idea and avoiding feature repetition. Repetitive characteristics, like irrelevant features, influence the speed and accuracy of learning algorithms and should be deleted, according to empirical findings from the features selection literature. As a result, pure relevance-based feature weighting methods fall short when it comes to feature selection for high-dimensional data with numerous duplicated characteristics.

In the feature selection process, the following elements must be considered: 1. The starting point, 2. The search strategy, 3. The subset evaluation, and 4. The stopping criteria Table 1 shows a comparative comparison of feature selection strategies based on these factors. We defined feature selection techniques to provide a summary of comparative analysis of search organization, feature generation, and evaluation measure that each feature selection technique entails, which can help practitioners choose a technique best suited to their goals and resources. In the study, nine feature selection approaches were examined. Mutual information (MI) of two random variables is used in mRMR (Minimal Redundancy and Maximal Relevance). MI is a metric that quantifies how dependent the two variables are on each other.

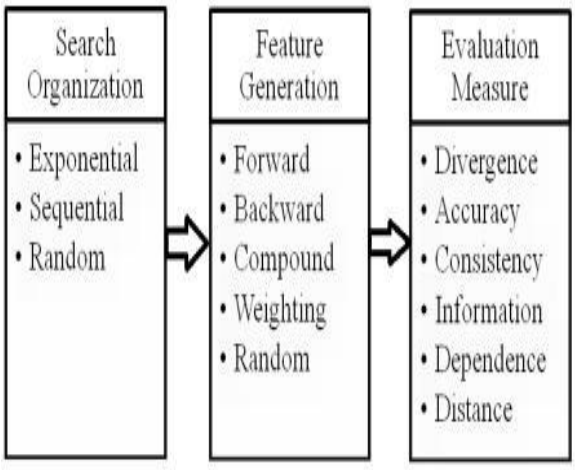

Fig. 4: Charachertization of Feature Selection Algorithms

Methods	Individual / Subset Feature	Starting Point	Search Strategy	Subset Generation	Subset Evaluation	Stopping Criteria	Used to Eliminate
Correlation Coefficient	Individual	Random Number of Features	Sequential	Forward Selection	Divergence (variance)	Ranking	Irrelevant Features
BW-ration	Individual	Full Feature Set	Sequential	-	Divergence (variance)	Ranking	Irrelevant Features
PAM	Individual	Random Number of Features	Sequential	Weighted	Distance/Information	Ranking	Irrelevant Features
mRmR	Subset	Random Number of Features	Random	Forward Selection	Mutual dependence/information	Ranking	Redundant Features / Irrelevant Features
I-RELIEF	Subset	Random Number of Features	Random	Weighted	Distance	Ranking	Irrelevant Features
CMIM	Subset	Full Feature Set	Sequential	Forward Selection	Conditional Mutual Information /	Relevance	Irrelevant Features
INTERACT	Subset	Full Feature Set	Sequential	Backward Elimination	Consistency	Relevance	Irrelevant Features
Genetic Algorithm	Subset	Full Feature Set	Random	Weighted	Consistency (cosine)	Ranking	Redundant Features / Noise
SVM-REF	Subset	Full Feature Set	Sequential	Backward Elimination/Weighted	Information	Ranking	Irrelevant Feature

Table 1: Charachertization Analysis of Feature Selection Algorithms

H. Feature Extraction/Transformation:

Feature extraction entails altering the original features to create more significant features. Feature extraction reduces the dimensionality of the selected characteristics by creating new variables from a mixture of others. In this context, feature extraction may decrease complexity and provide a straightforward data representation. Each variable in the feature space is represented as a linear combination of the original input variable. Karl's Principle Component Analysis (PCA) is the most popular feature extraction method. PCA has been proposed in a variety of forms. PCA is a non-parametric approach for extracting the most important facts from a group of redundant or noisy data.PCA is a linear data transformation that minimizes redundancy (as measured by covariance) while maximizing information (measured through the variance).

ISSN No:-2456-2165

The effects of various dimensionality reduction methods (including feature subset selection using information gain (IG) and wrapper methods and feature extraction with different flavors of PCA methods) on classification performance have been empirically tested on two different types of data sets to investigate the relationship between these methods and their effects on classification performance (email data and drug discovery data). The results reveal that PCA feature extraction (transformation) significantly depends on the data.

IV. FEATURE EXTRACTION/TRANSFORMATION METHODS

It is critical for further data analysis; whether it is pattern recognition, de-noising, data compression, visualization, or anything else, the data must be represented in a way that makes analysis easier. To discover an appropriate transformation, many basic approaches have been devised. Independent Component Analysis (ICA) is a linear transformation approach in which the intended representation is one in which the statistical dependency of the representation's components is minimized. The use of ICA for feature extraction is driven by neuroscience findings that imply a similar concept of redundancy reduction may explain some parts of the brain's early sensory input processing. Like the closely related approach of projection pursuit, ICA has applications in exploratory data analysis. The principle of redundancy reduction motivates the usage of feature extraction. ICA algorithms are divided into two groups. Some algorithms are based on mutual information reduction, while others are based on non-gaussianity maximization. Mutual information may be defined as a reduction in uncertainty about variable X due to the observation of variable Y. As a result, we are looking for maximally independent components using an algorithm that aims to reduce mutual information. Focusing on non-gaussianity is another technique to evaluate the independent component. One method for extracting the components is to make each one as far away from the normal distribution as feasible. In most cases, five requirements must be followed to execute ICA: 1 – the source signals must be statistically independent; 2 – the number of source signals must equal the number of mixed observed signals, and mixtures must be linearly independent of each other; 3 – the model must be noise free; 4 – data must be centered; 5 – the source signals must not have a Gaussian probability density function (pdf), except one signal source that can be Gaussian.

An orthogonal transformation is used in Principal Component Analysis (PCA) to turn samples from correlated variables into samples with linearly uncorrelated features. Principal components are new characteristics that are fewer or equal to the starting variables. Because PCA is an unsupervised approach, it does not incorporate data label information. When data is properly distributed, primary components are self-contained. PCA is a basic nonparametric approach for extracting the most important information from a group of redundant or noisy data. This is the fundamental rationale for its use.

By removing the final principle components that do not contribute significantly to the observed variability, PCA minimizes the number of original variables. PCA is a linear data transformation that reduces duplication (measured by covariance) while increasing information (Measured through variance). Principal components (PC) are new variables with two properties: 1) each PC is a linear combination of the original variables, and 2) the PCs are uncorrelated to one another, removing unnecessary information [12]. Data compression, image analysis, visualization, pattern identification, regression, and time series prediction are some of the most common PCA applications. PCA has certain drawbacks. For example, it presupposes that the connections between variables are linear.2) Its meaning is only comprehensible if all variables are considered to be numerically scaled. 3) It lacks a probabilistic model framework, critical in many situations like mixture modeling and Bayesian decision-making

Methods	Time Complexity	Data Type	Application
mRmR	-	Discrete / Continues	Microarray Gene Expressions
I-RELIEF	O(Poly(N))	Discrete / Continues / Nominal	Protein Folding and Weather Prediction
CMIM	$O(N^3)$	Boolean	Image Classification
Correlation Coefficient	-	Discrete / Continues	-
BW-ration		Discrete / Continues	-
INTERACT	$O(N^2M)$	-	-
Genetic Algorithm	-	-	Pattern Recognition, Machine Learning, Neural Networks, Combinatorial Optimization, Hyper Spectral Data
SVM-REF	-	Discrete / Continues	Microarray Gene Expressions
PAM	-	-	Microarray Gene Expressions

Table 2: Comparison of FSA's

V. DISCUSSION

Because of the noisy data, feature extraction approaches were far more suited for the automated identification of ophthalmology illnesses than feature selection methods. Because the majority of biological datasets contain noisy data rather than useless or redundant data. Feature selection is a tool that may be used to remove unnecessary and/or superfluous features in various applications. There is no unique strategy for selecting features that can be used in all applications. Some strategies were utilized to remove unimportant characteristics while avoiding duplicate ones. Purely relevance-based feature weighting methods are inadequate for feature selection. Subset search algorithms look for potential feature subsets based on an evaluation metric that measures how excellent each subset is.The consistency and correlation measures are two current assessment tools that have been proven successful in deleting irrelevant and duplicate characteristics. Experiments demonstrate that the number of iterations necessary to discover the optimum feature subset is usually at least quadratic to the number of features. As a result, existing subset search methods with quadratic or

greater time complexity in dimensionality do not have adequate scalability to cope with high dimensional data. Filters and wrappers are two types of feature selection strategies. Because the feature selection procedure is tuned for the classification algorithm, wrapper approaches often outperform filter methods.However, if the number of features is huge, they are usually far too expensive to employ because each feature set must be assessed with the trained classifier.

VI. CONCLUSION

It is planned to survey feature selection and extraction. Both strategies have the same goal: to reduce feature space to better data analysis. This aspect becomes even more significant when dealing with real-world datasets, which might contain hundreds or thousands of characteristics. The main difference between feature selection and extraction is that the former reduces dimensionality by selecting a subset of features without transforming them, whereas the latter reduces dimensionality by computing a transformation of the original features to produce other, more significant features. Table 2 shows traditional approaches, their subsequent advancements, and some novel applications for feature selection.Feature selection increases understanding of the process under examination by highlighting the characteristics that have the greatest impact on the phenomena under discussion. Furthermore, the computation speed and accuracy of the chosen learning machine must be evaluated since they are critical in machine and data mining applications.

REFERENCES

[1.] N. Chumerin and V. Hulle, M. M, "Comparison of Two Feature Extraction Methods Based on Maximization of Mutual Information" In: Proceedings of the 16th IEEE Signal Processing Society Workshop on Machine Learning for Signal Processing, pp. 343–348, 2006.

[2.] H. Motoda and H. Liu, "Feature selection, extraction and construction" In: Towards the Foundation of Data Mining Workshop, Sixth PacificAsia Conference on Knowledge Discovery and Data Mining (PAKDD 2002), Taipei, Taiwan, pp. 67–72, 2002.

[3.] L. Ladla and T. Deepa, " Feature Selection Methods And Algorithms", International Journal on Computer Science and Engineering (IJCSE), vol.3(5), pp. 1787-1797, 2011.

[4.] G. K. Janecek and G. F. Gansterer et al, "On the Relationship between Feature Selection and Classification Accuracy", In: Proceeding of New Challenges for Feature Selection, pp. 40-105, 2008.

[5.] M. Dash and h. Liu, "Feature Selection for Classification", Intelligent Data Analysis, vol. 1, pp. 131-156, 1997.

[6.] P. Soliz et al, "Independent Component Analysis for Vision-inspired Classification of Retinal Images with Age-related Macular Degeneration", In: proceeding of IEEE Int'l. Conference on image processing SSIAI, pp. 65-68, 2008.

[7.] Deka and K. Kumar Sarma, "SVD and PCA Feature for ANN Based Detection of Diabetes Using Retinopathy", In: Proceedings of the CUBE International Information Technology Conference, pp.38-44, 2012.

[8.] Y. Zheng et al, "An Automated Drusen Detection System for Classifying Age-Related Macular Degeneration with Color Fundus Photographs", In: IEEE 10th International Symposium on Biomedical Imaging, pp.1440-1443, 2013.

[9.] S. Rajarajeswari and K. Somasundaram, " An Empirical Study of Feature Selection For Data Classification", International Journal of Advanced Computer Research, vol.2(3) issue-5, pp. 111-115, 2012.

[10.] Veerabhadrappa, L. Rangarajan, "Bi-level Dimensionality Reduction Methods Using Feature Selection and Feature Extraction", International Journal of Computer Applications, vol. 4(2), pp. 33-38, 2010.

[11.] L. Yu, and H. Liu, "Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution", In: Proceeding of the

[12.] S. Cateni, et al, "Variable Selection and Feature Extraction through Artificial Intelligence Techniques", Multivariate Analysis in Management, Engineering and the Science, chapter 6, pp.103-118, 2012.

[13.] T. Howley and M. G. Madden et al, " The Effect of Principal Components Analysis on Machine Learning Accuracy with High Dimensional Spectral Data" Knowledge Based Systems, vol. 19(5), pp. 363-370, 2006.

[14.] Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection" Journal of Machine Learning Research, vol.3, pp. 1157-1182, 2003.

[15.] Veerabhadrappa, L. Rangarajan, "Multilevel Dimensionality Reduction Methods Using Feature Selection and Feature Extraction", International Journal of Artificial Intelligence and Applications, vol. 1(4), pp. 54-58, 2.