International Journal of Computational Intelligence Research
ISSN 0973-1873 Volume 13, Number 3 (2017), pp. 403-424

http://www.ripublication.com

A comparative Study of Machine Learning Classifiers over Gene expressions towards Cardio Vascular Diseases Prediction

E. Neelima
Assistant Professor, Department of CSE
GITAM University, Visakhapatnam, AP, India.

M.S.Prasad Babu
Professor & Vice-Principal, Department of CS&SE
Andhra University, Visakhapatnam, AP, India.

Abstract

Contemporary solutions in genetic technologies like the NGS and GWS are playing a vital role in identifying the complex disease genetic mapping, and in the process of evaluating the genetic structure for complex diseases, machine learning methods are turning out to be a potential solution. In the implementation of machine learning methods, feature selection models are very important in attaining optimal results and accuracy from the machine learning models. In this paper, emphasis is on getting insights in to the varied kinds of machine learning methods, feature selection models that could make significant difference in terms of improving the quality and outcome for the process. An explorative study of varied feature learning models and the performance analysis of feature selection models for machine learning algorithms were carried out, and it is evident that despite of the efficacy shown by certain models, still there is considerable scope for improvement.

Keywords: Micro Array, gene expression data, gene expression profiling, Machine Learning. Classification, feature selection, IG, GS

INTRODUCTION

Disease phenotypes are usually genetically complex and due to the combination of genetic variation in resulting in varied loci. Major challenge that is envisaged in the medical genetics is about determining set of genetic makers when combined with conventional risk factors that could be resourceful in individual's susceptibility for observing varied complex disorders. Advancements that were taking place in the genetic technologies like the NGS (Next-generation sequencing) and GWS (Genome Wide Association) has revolutionized the way in-depth analysis is carried out in variations in human genome. NGS and GWS are supporting in investigation of genetic architecture of complex diseases, and in developing risk prediction models to personalize prevention and treatment alternatives for varied diseases [1].

GWA studies has been very successful in identifying numerous range of genetic variants which are associated to complex human diseases and other such characters [2]. Predominantly the models has relied upon statistical association testing approach, in which only a small portion of heritability are observed, and such effects are not predictive enough for clinical utility, raising questions on the scope of observing "missing heritability" [3].

Considering the nature of multifaceted disease complexities, the contemporary research efforts focusing on concept of relations among the position of gene on chromosome, and the phenomenon of epistasis (effect of one gene (locus) being dependent on the presence of one or more 'modifier genes'), which is a major factor of missing heritability [4]. Moreover the current testing strategies are able to focus on single variant associations of genes, hence not fitto address the contemporary requirements.

Gene relations that are influenced from the phenomenon of epistasis interactions profoundly are distinct to the testing procedures of conventional single-variant association [5]. Domain experts are of view that, it is very essential to focus on "one variant at a time" rather than focusing on complete set of variants. Network-centric approaches could be more resourceful options in understanding and decoding complexity of relationships of genotype-phenotype, as they are characterized by interactions of gene-gene and gene-environment [6] [7]. Despite the fact that traditional models of statistical oriented testing procedures resulted in effective identification of numerous susceptible loci, still, if the potential risk variants comprised in gray zone are ignored, there are significant chances that it might impact heritability variation [1] [8].

Considering such limitations, in this paper, our focus is upon machine learning approachesthat relies on hidden interactions amidst the genetic panels and numerous other risk factors which could impact disease risk in individuals. The scope of the review carried out in this manuscript is exploring the genetic feature selection measures and adapting network-guided disease scope predictive models and divergent learning models inculcated by those features for optimal prediction of phenotypic response variables (quantitative phenotypes in regress problems or class labels with case-control classification).Computational approaches that relies on epistatic based machine learning models could offer effective framework for using the complete spectrum of genetic information, whilst predicting an individual's risk of proneness to a disease. Though such methods could be resourceful, still they are at nascent phase. Implementing solutions for genetic feature selection, by adapting computational algorithms that are robust and scalable, is very essential for effective data mining from current GWA studies [9].

Aiming at developing quality solutions and evolving computational and scalable computational solutions, scope for contemporary solutions development, review of the existing models, evaluating the pros and cons are carried out in this article. Developments in the recent past, pertaining to developing accurate and robust predictive models (machine learning oriented) are detailed in section-2 of this paper. In section-3, model validation approaches are discussed, and the section -4 reflects the network level analysis performance of genetic variants and the assessment of current state of data mining solutions. Section-5 offers inputs on some of the existing limitations and scope of development towards gaining insights in to individual predisposition for genetically complex diseases.

MACHINE LEARNING STRATEGIES

Machine learning methods has been very resourceful for interpretation of huge genomic datasets and are used for annotating vivid range of genomic sequence elements [10] For instance, using the machine learning methods for learning the pattern of recognizing the position of TSSs(Transcription Start Sites)for a genome sequence is an effective solution [11]. Machine learning algorithms can be trained for varied elements like positioning of nucleosomes [12], promoters [13], splice sites [14], enhancers [15].

If sequence elements list for a specific type is focused upon, training the machine learning solutions for such elements becomes much easier [16]. As the input data can be used by the algorithms for generating other genomic essays, the task gets much simpler. . Gene expression data shall be used for learning and distinguishing varied disease phenotypes and also shall be able to identify disease biomarkers.

Chromatin data could be very useful for annotating the genome in unsupervised manner, and hence it increases the scope of identifying the functional elements of new classes. Also, the machine learning models can be very resourceful in assigning functional annotations to the genes and usually takes the form of Gene Ontology term relationship [17]. As an alternative for Gene Ontology term prediction, some solutions even focus on co-functional relationships which can adapt machine learning method output from a network, and the genes are denoted as nodes. Edges amidst two genes indicate the common functionality [18].

Vivid range of machine learning algorithms is evident from review of literature for understanding the mechanism underlying the gene expressions. While some techniques target predicting the gene expression, by relying upon DNA sequence [19], few of the models consider histone modification based ChIP-Sequence [20]. Some of the models have also focused on transcription factor binding [21] at gene promoter region.

Some of the novel solutions attempt to focus both on the model of expression of all genes in a cell and also training the network model [22]. Similar to a co-functional network, every node of a gene expression network indicated gene, however in certaincases edges indicate the regulatory relationships among the transcription factors and related targets.

Majority of the aforesaid issues are addressed using statistical mapping approach and there is thin distinction between the machine learning and statistics [23]. In the following sub-section, scope and implementation of machine learning applications in genomics and genetic solutions were considered. Some of the key methods of machine learning and prime factors taken in to account whilst applying such methods to genomics are taken in to consideration. Also, an overview of varied type of research elements for which the machine learning models are appropriate are also considered and elaborated discussion of machine learning applied to specific subdomain of genetics and genomics as applied elsewhere [24].

- ***Phases of machine learning***

Machine learning methods take place in three stages. Firstly an algorithm that can lead to successful learning has to be developed. Secondarily, the algorithm has to offer large collection of both TSS sequences and non TSSs sequences [25] (see fig.1). Training set of DNA sequences is provided as input towards learning procedure, wherein the binary labels denote the status of each sequence as centered over a TSS or not. Learning algorithm generate a model that are used along with a prediction algorithm. The gradient of red-blue in the fig.1 denotes the scores of varied motif models against to a DNA sequence.

Labelled sequences are processed by the algorithm and it also stores the model. In the final stage, new set of unlabeled sequences are provided to the algorithm, and it shall

use the solution for predicting the labels for the sequence. In the instance of successful learning, majority of the predicted labels could be right.

The process of developing the algorithm design, ensuring the training of algorithm and testing the process is an effective ways for building a robust system. For instance, the design –learn-test method provides is a sound way of hypothesis testing of the solution. Secondarily, the resultant theory if handled appropriately can be deployed as models.

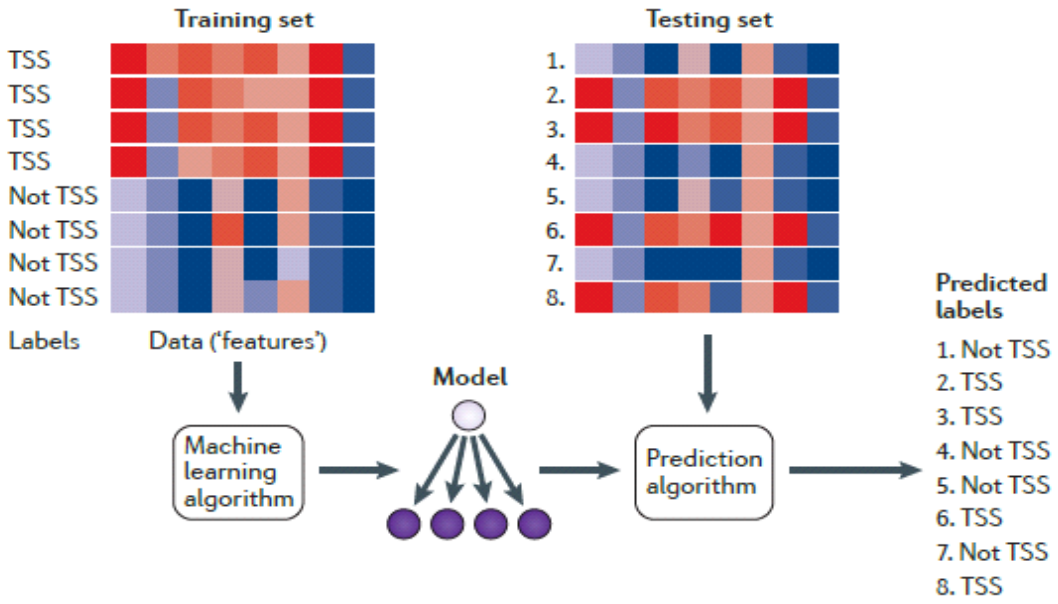


Figure 1: The process of machine learning on Genomes [25]

- ***Classification of Machine Learning Models***

Machine learning methods can be classified in to three major categories as supervised, unsupervised and semi-supervised models of learning. In the supervised models, the model is trained on categorized examples and accordingly used for making predictions for unlabeled examples. In the unsupervised methods, the model observes dataset without using labels. An illustration of the supervised and unsupervised machine learning models of a gene-finding algorithm is depicted in Fig.2. The prototype reflects the fundamental properties adapted in protein-coding gene. Model adapts takes DNA sequence of chromosome as input, and it generates detailed gene annotations as output. Inability towards observing any overlapping genes or the resulting isoforms of same gene in the untranslated region. In the illustration provided, the trained model can use the learned properties for identifying the additional genes which look similar to the genes in the training set.

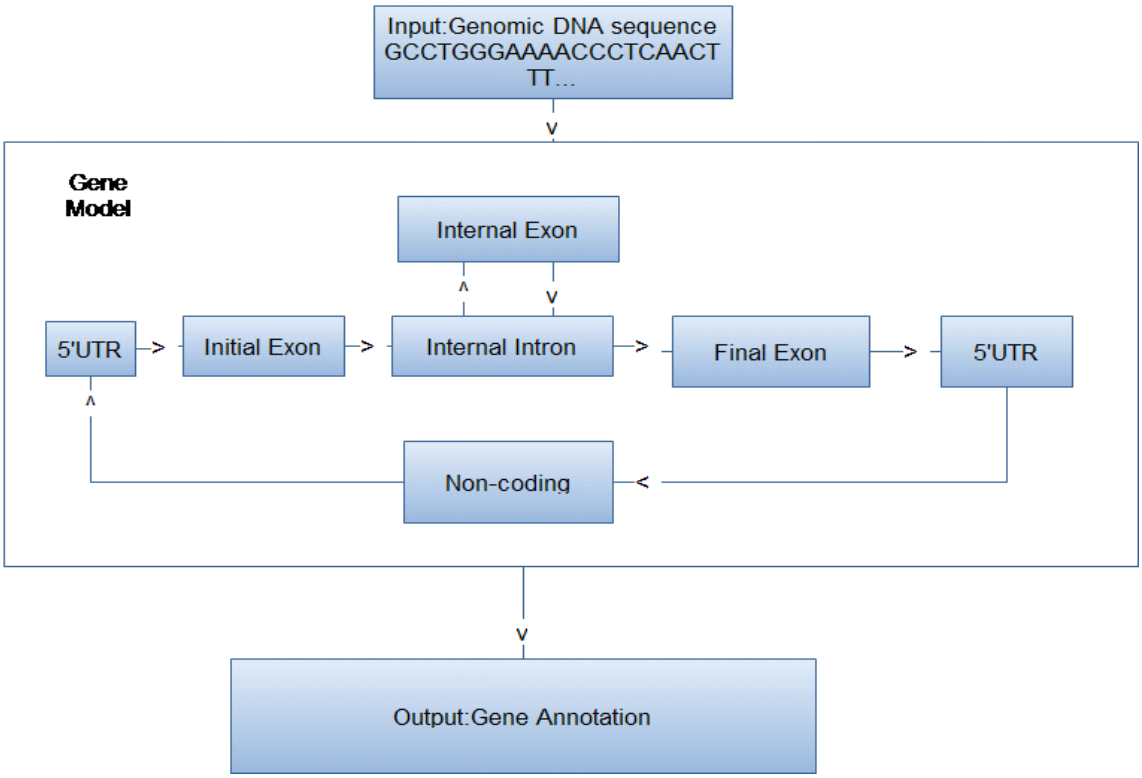


Figure 2: Prototype of Gene Finding Process

Unsupervised learning is very vital in the absence of unavailability of any kind of training sets. In the instances where the training set do not exist, unsupervised learning is essential. In such conditions, the machine learning algorithm use the unlabeled data and requisite volume of varied labels for assigning the input[26]. Also, it automatically partitions the genome in to segments, and assigns a label for every segment, targeting at assigning the same label to segments comprisingrelated data. The unsupervised approach needs an addedprocess of analyzing semantics for every label by a manual process However, the scope for enabling training in the instances of unavailability of labelled examples is a constraint. Also, the ability towards understanding potentially the novel type of genomic elements are also few of the key factors to be considered in the process.

Semi-supervised model of machine learning is the intermediary process of supervised and unsupervised model [27]. In the semi-supervised approach, set of data points are received by the algorithm, and only a subset or fraction of such data are the labelled ones. The learning procedures are initiated with development of a initial gene-finding model and relies on training data with labelled subset.

In the further process, the model scans the genome and assigning of tentative labels for the genome takes place. Tentative labels are used for improving the learning model, and until the new genes are identified, the procedure is iterated.

The semi-supervised approach can be very resourceful than a fully supervised model as the model is capable of shaping of much larger set of genes.

When supervised learning models are feasible, additional set of unlabeled data points can be easily obtained. In such conditions both supervised or semi-supervised can be considered. In the semi-supervised learning, having certain assumptions over the data set is essential [27] and generating such data could be very complex. Hence, focusing on semi-supervised learning can be resourceful when there is small volume of labelled data and huge volume of unlabeled data.

Application of machine learning models focus on two key elements like prediction and interpretation. In the case scenario of predicting based on ChIP-seq experiment, locations for given transcription factor shall impasse to genomic DNA. Task is analogous to the TSS prediction task except to the fact that labels were derivative of ChIP-seq.

Differentiation amidst the discriminative and generative models certainly plays a major role in the exchange of Interpretability and performance. In the case of a generative approach, emphasis is on building a full model of features distribution for every two classes and it is compared to how the distribution differ from one level to the other. But in the discriminative model, the system focus on accurately modeling only the boundary amid of two classes. In anintrinsic view, it can be stated that the discriminative approach only focus on modeling the conditional distribution of the model and not on combined distribution of labels and features.

SVM (Support Vector Machine) is an effective model of discriminative algorithm [28] has objective of learning to output a value of 1 in the instance of a positive training lead and value of -1 in the instance of negative training is offered.

The key benefit of discriminative modeling approach is that it supports in achieving better performance than the generative modeling approach even in the instances of unlimited training data [29]. In preparation, analogous generative and discriminative approaches repeatedlycongregate same solution. Performance of generative approaches can be more effective even with limited set of training data.

If the quantum of labelled set of training data is large, in such instances, discriminative approach shall be effective in terms of predicting desired outcome more accurately under test conditions of previously unseen data. The SVM model illustrates the accuracy of training examples over PFSM model. It is imperative from the model that the discriminative approach shall tend to provide estimation of accurate ones.

One of the key issues and the negative elementsuch accuracy is that by focusing on a single problem well, the other problems are ignored in the case of discriminative modeling. In the generative model, various factors could be considered, but in the case of discriminative model, only answer to a single question for which it is

considered shall be generated. Hence, there is always a trade-off between generative and discriminative approach.

Hence, selecting a model between the generative and discriminative approach could lead to traction of either accuracy or interpretability of the model. Despite of distinction between the models playing a vital role in interpretability of the model, even the number of parameters used could also be important. Complexity of a model could be impacted by the way a simple model is chosen or by adapting a feature selection strategy for restricting the complexities in the model chosen.

- ***Significance of Learning Efficacy***

In machine learning models, outcome is based on effectiveness of how accurate the process of encoding the earlier knowledge for the issues in hand. Also, it is very important that the practical application of the model constitute gaining insights in to problem specifics, choosing right kind of algorithmic approach and accurately encoding. Optimal machine learning based algorithms can be resourceful [30] as it can match the earlier knowledge of the problem and can lead to successful analysis.

Prior knowledge is determined set of discreet data offered as key input to the machine learning algorithm. Prior knowledge shall be implicitly encoded to the learning algorithm and certain type of solutions are chosen over the other options [31]. Range of input data sets, referring to data and any kind of pre-processing shall be guided by prior knowledge of the application and the data used.

In probabilistic framework, certain categories of prior knowledge are mentioned overtly by detailing a prior distribution over the data. Uniform prior is a way of common prior distribution, and can be resourceful in many contexts. The efficacy of the solution is that despite of sequences comprising nucleotides are ignored in a position, still it can be assigned a non-zero probability by the model.

Prior information is certainly a non-probabilistic models, and combination of prior knowledge in to non-probabilistic methods could lead to more implications. However, one class of discriminative methods offer a generic mechanism for demonstrating prior knowledge and if it can be encoded in to generalized notion of similarity [32]. Formal probabilistic priors can be adapted in aggregation with any kernel methods [33].

- ***Feature selection***

Researchers have to focus on the kind of data to be provided as input, whilst planning for application of machine learning methods. It is very essential to have classification and insights on what kind of data have to be relevant to the application, for training the algorithm.[34].Suchclassifiers shall be resourceful in two ways, firstly, the classifier shall establish accurate diagnosis even in a typical demonstration or histopathology. Secondarily, the model developed from the learning phase shall carry

out feature selection, subsets of genes identification along with expression of patterns that support in addressing varied aspects of problem.

Feature selection methods shall be implemented with supervised learning algorithm models, wherein the algorithm is provided large set of features and the system takes a decision towards ignoring some or many features, automatically, on the basis of subset of features that suits the tasks to be performed. If the classifiers are accurate enough, it might be result in an inexpensive clinical assay[35] and it is very essential to ensure to train the most accurate possible classifier [36]. Due to the conditions where high-dimensional datasets like the genomic, epigenomic, proteomic and metabolmic datasets envisage issues from the curse of dimensionality [37]. such phenomenon leads to quality performance in terms of training data. Feature selection models and dimensionality related techniques like the principal component analysis or multidimensional scaling focus on addressing the problem by projecting the data in descending order as higher to lower dimensions.

MACHINE LEARNING AND USE OF GENETIC RISK FACTORS

Numerous researchers have focused on the usage of machine learning methods in the genome-wide data on genetic variants [38] when compared to the other kind of machine learning studies on other types of genomic datasets like the gene expression profiles.

Also, the combinative approach of advanced feature selection algorithms and predictive modeling can be deployed even to restricted set of studies, despite of studied despite of the positive results that are attained from the models. [6] [39. It is imperative from the review of many feature selection models that the feature selection models can be very result oriented in results prediction [8] [39] [40] [41]. But it could be very challenging to obtain predictive signals from high-dimensional datasets that are usually gathered from GWA or NGS studies. The following subsections describe some of the feature selection category that are in existence and has been resourceful in improving the machine learning outcome.

- ***Filters***

Filter methods for genetic feature selection are very common in GWA studies because of the easeof implementation, human interpretability of the results and the lower levels of computational complexity. In basic models too, the filter methods compute a univariate test statistic individually for every genetic feature and ranking of features based on statistical values observed is carried out. Accordingly, the highest ranked features are chosen to develop a final set of features chosen, and used for training the predictive model. The number of features chosen for a model might be either pre-determined or considered based on the significance of threshold defined for every test

statistic. Varied range of statistical tests have been used in GWA studies, like the popular models like Armitage trend test and Fisher's exact test[42].

Increasing number of statistical approached are generated for rare variants and NGS data [43]. Contemporary methods of filtering could be used for selecting specific risk variant combinations pertaining to a disease risk.

- ***Wrappers***

Wrappers comprise three key components, wherein the first component is the search algorithm adapted for methodically negotiating the space for possible feature sets. The second component is the scoring function that handles the process of predictive accuracy weighing for chosen feature subsets. The third element is the learning algorithm for supporting feature selection procedure [44].

Despite the fact that there are many methods of scoring used along with wrapper methods for predicting error evaluation in a training set, certainly the method of cross-validation error turns out to be a significant solution. Also, the feature selection models can be used for wrapping any learning models and it could lead to more optimal results, upon training the method efficiently [39].

- ***Embedding***

In Embedding methods, the feature selection mechanism in-built to its training algorithm [45], whereas in the predictive models, it relies completely on a subset of the original features. LASSO (Least Absolute Shrinkage and Selection Operator) is one of the reputed embedded method which is vividly used in GWA studies [38] [41], [46]. While some machine learning approaches support scaling up for genome-wide level, in LASSO such scope has been introduced by some of the novel training algorithms like the coordinate descent methods that possess high computational efficiency.

Wrappers and embedded methods comprise the characteristic to generate better results than filter methods in varied applications [41] [39] [47]. However, the challenge is that if the implementation is not appropriate, models might envisage failure in prediction beyond the trained data and could lose the purpose of being a prediction model.

- ***Complex phenotypes***

Among the key challenges of feature selection, accurate estimation of the prediction performance of machine learning models on new samples that are not envisaged during the training phase is a major issue. Considering the facts like the immense dimensionality of contemporary GWAS and NGS studies, it might be very

challenging to identify genetic features that could be apt to training set, however, it might fail in terms of generalizing the invisible data, a phenomenon considered as model overfitting. In many of the applications of genomic predictors, varied instances of the professed selection bias, leading to cross-validation is used for performance estimation of learning algorithm only, but not limited to primary feature selection that is performed on whole data. Hence, it leads to information pilferage and grossly over-optimistic results. [48The aforesaid two-level technique is termed asnested cross-validation [49].

From the review of the models and the illustration of such models, it is imperative that, for actuallycorroborating the generalizability of predictive risk models, using large volume of datasets comprising independent sample without any overlap amid of examined cohorts [50] are a potential solution. But even such decisions has to be taken in accordance to the requirement as to whether the emphasis is on predictive model itself or predictive variants that shall be chosen by the model [51].

Despite of numerous comparisons on various feature selection methods and frameworks comprising predictive modeling for individual cohorts [39] [40] are carried out, hardly any definitive results pertaining to optimal results to be obtained by applying any one specific method.

In the instances where epistasis interactions are involved, optimal results may not be feasible by deploying only single locus filters and wrapper methods could envisage computational issues in the case of large volume datasets, upon combining to complex prediction models to the system.

Usage of molecular networks as a previous information in developing the predictive models, despite the absence of single-locus marginal effects, itmay turn out to be synergizing and could lead to disease phenotype upon combining with other elements. Hence it can be stated as a popular model for mapping genetic loci identified in GWA or NGS studies, for establishing biological pathways to depict the probable cellular mechanisms for observed genetic and phenotypic variation. [52].

Machine learning based predictive models developed on gene expression outlining has reflected the benefits of choosing pathway activities as features for enhancing classification accuracy levels, than the models that merely target individual gene expression levels. [53]. It is also imperative that the established context of GWA datasets which the pathway analysis could offer both the mechanistic insights and also improve discrimination power based on customized statistical data mining techniques like the HyperLasso.

Models consider the pairwise connections in a way of defining classic definition of epistasis that involves single and double-deletion experiments in model organisms. In the instance of performing computationally efficient exploration of genetic interactions, a posteriori finding and heuristic search schemes are not able to assure

about the pairs of genetic risk factors that are detected and ultimately be the key ones for enhanced predictive power between possible variant combinations. [54].

It is imperative from the review that combination of individual level gene expression estimations with background networks, amidst of transcription factors signify the possibility for identifying and making use of disease specific sub network for reducing the false quantum of false positives and false negatives emerging as outcome of technical variability and genetic heterogeneity, and also in terms of improving the prediction at individual levels in clinical outcomes like the cancer metastasis or survival time [55].

In one of the contemporary solutions, a principled method which use the genetic algorithm guided by a structure of chosen gene interaction network for discovering small groups of connected variants that are combined with disease outcome on a genome-wide scale for estimating outcome [56].

PERFORMANCE ANALYSIS

In a performance analysis of the machine learning technique on a dataset that was preprocessed, and is converted to Attribute-Relation File Format. As a furtherance, IG and GS feature selection techniques were applied on the dataset and feature set of selective set were obtained.

Each feature selection techniques could support with a unique reduced set of features. Every such distinctive feature set were tested for accuracy using 6 different set of classifiers like Logistic Regression (LR), NB, SVM ,DT, Multi-Layer Perceptron (MLP) and Adaboost. Similar kind of corroboration is carried out on common set of attributes generated from feature selection across all the datasets. In the final step, validation and analysis of the minimized feature sets are performed depending on anatomic relevance. Figure 3 indicates the figurative representation of the approach.

GS is used for introducing the principle of evolution and genetics to the chosen dataset of problem. Genetic search algorithm shall support in mimicking the human behavior. GS approach is initiated by taking in to account the sample of attributes as chromosomes, and considering best of such chromosomes for feature selection. For evaluation of individual cases and process of cross over, fitness function is derived and the mutation were followed.

The process has been implemented on numerous generations and the results were analyzed [57]. IG is used for evaluating the credibility of an attribute by determining information gain with respect to class [58]. Using the ranker method for filtering the attributes and then ranked them for selection. By engaging inrepetitive occurrence ofiterations like changing the ranking parameters, reduced attributes set are obtained.

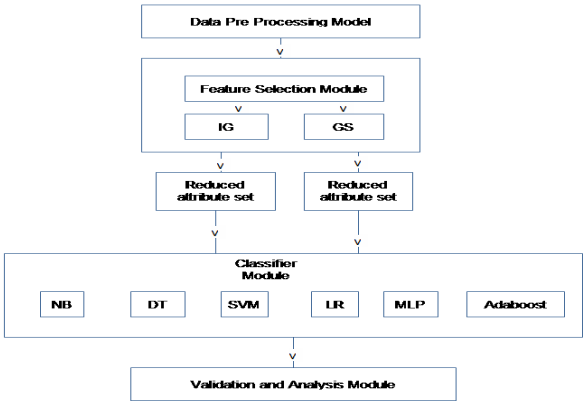


Figure 3: Schematic Diagram of Proposed Approach

Considering the acceptance and effectiveness, an exhaustive list of classifiers were selected for validating the set of features. The classifiers that were opted for the analysis are NB, DT, LR, MLP, SVM and Adaboost [59]. The instance of implementing the classifiers, a major dip in the performance is envisaged when there are outliers in the datasets.

- ***Data Statistics***

Cleveland dataset, Hungarian dataset, Switzerland dataset and Long Beach dataset were the four datasets that are considered for analysis. All the datasets were preprocessed and number of instances available for each of the dataset is depicted in table.1. While the Cleveland and Hungarian datasets possess uniform distribution, Long Beach and Switzerland datasets were skewed.

TABLE 1: DATA STATISTICS FOR EACH DATA SET

Data Set	Total # of Instances	Class-0	Class-1	Class-2	Class-3	Class-4
Hungarian	294	188	37	26	28	15
Switzerland	123	8	48	32	30	5
Cleveland	283	157	50	31	32	11
Long Beach	200	51	56	41	42	10

One of the intrinsic observation in the datasets are that, class 2 and class 3 have comparable kind of distribution of the number of instances and class 4 comprising least number of instances. Preprocessed data is provided to Weka tool for analysis.

In a usual procedure for diagnosis of cardiovascular disease, three stages are involved. Firstly, the test for risk factors evaluation followed by stress testing and the third stage of coronary angiography [60]. 75 attributes in the disease dataset are related to the aforesaid procedures as mentioned in UCI repository. The analysis is carried out in

two stages,wherein in first stage only 13 commonly used attributes and 5 classes were evaluated and then using 75 attributes and 5 classes were used for evaluation.

- ***Feature selection and performance statistics***

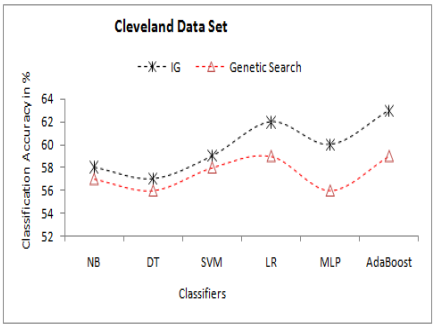
The feature selection techniques baseline,IG and Genetic Searchthat were chosen and the list of reduced set of attributes are depicted in Table 2. Once the GS is applied, the datasets of both LongBeach and Switzerland datasets have same attribute exang and hence it can be stated that feature exang is considered as a noticeable feature among the 13 key features. In furtherance, reduced features are validated using 6 classifiers and the results are depicted in Figure 4.

TABLE 2: ATTRIBUTES SHORTLISTED FROM 13 KEY ATTRIBUTES, USING GS AND IG

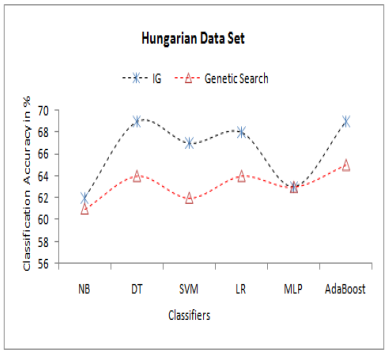
	Hungarian Data Set	**Switzerland Data Set**	**Cleveland Data Set**	**Long Beach Data Set**
GS	SEX CP EXANG SLOPE (Features Count: 4)	EXANG (Features Count: 1)	OLDPEAK SEX THAL CA EXANG CP THALACH EXANG SLOPE (Features Count: 8)	EXANG (Features Count: 1)
IG	SLOPE EXANG CP OLDPEAK THALACH SEX TRESTBPS THAL (Features Count: 8)	THAL FBS TRESTBPS CP (Features Count: 4)	THAL CP CA EXANG OLD PEAK (Features Count: 5)	THAL THALACH RESTECG OLDPEAK CHOL SEX FBS TRESTBPS CA CP SLOPE (Features Count: 11)

Validation is carried on the basis of rightly classified instances. And from the implementation of GS validation and IF feature selection technique, it is evident that

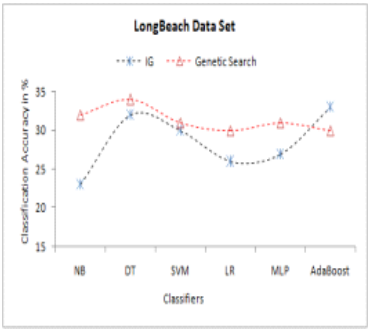
all the 13 features are significant and feature reductions are not being significant. Performance dip is observed for LR, and also for a specific dataset when SVM is implemented in GS. In overall, the results depict that though the feature selection techniques are applied, still the performance level improvements in overall performance of the classifiers have become a challenge, and thus there is need for focusing on residual features and their importance for effective decision making.



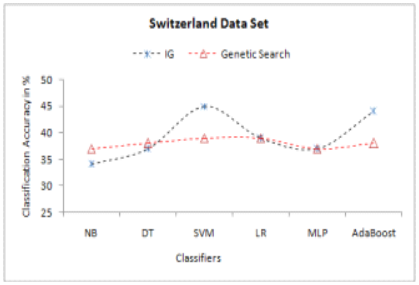
Classification Accuracy on Cleveland Dataset



Classification Accuracy on Hungarian Dataset



Classification Accuracy on LongBeach Dataset



Classification Accuracy on Switzerland Dataset

Figure 4: Classifier performance analysis on 13 optimal attributes

Further, the classifiers performance were assessed against all the 75 attributes (advocated by UCI) were considered, and the Table 3 depicts the names and numbers of attributes for each of the dataset, after the feature selection model was applied.

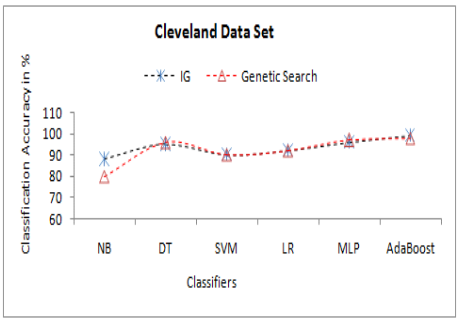
Results of applying the attributes selected for both the feature selection techniques and as per the stage-1, validation is carried out for classifiers and results were depicted in Figure 5.

TABLE 3: ATTRIBUTES SELECTED FROM 75 ATTRIBUTES BY ADAPTING IG AND GS

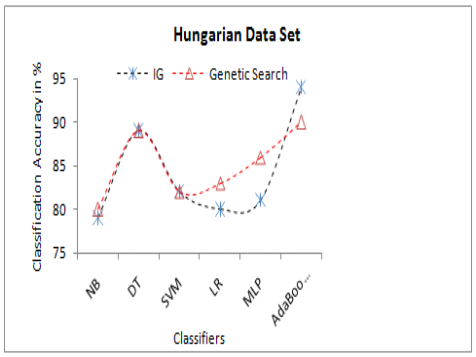
	IG	**GS**
Hungarian Data Set	OM1 EXANG LMT THALTIME RCAPROX CXMAIN LADPROX RELREST (RELIEVED AFTER TEST) RCADIST PAINEXER (PAIN DURING EXERTION) LVX4 SLOPE OLDPEAK LADDIST CP (Features Count: 15)	LADDIST SLOPE RCAPROX LMT CXMAIN LADPROX PAINEXE (Features Count: 7)
Cleveland Data Set	RCADIST LMT LADDIST OM1 RCAPROX THAL LADPROX CXMAIN (Features Count: 8)	LADPROX RCADIST LMT LADDIST CP CXMAIN,OM1 RCAPROX CA THAL OLDPEAK (Features Count: 11)
Long Beach Data Set	RCAPROX LMT CXMAIN LADPROX OM1 LVX4 LADDIST RCADIST (Features Count: 8)	RCAPROX LMT CXMAIN LADPROX OM1 LVX4 LADDIST RCADIST (Features Count: 8)
	LMT	LMT

Switzerland Data Set	RCAPROX RAMUX LADPROX CXMAIN OM1 LADDDIST (Features Count: 7)	RCAPROX RAMUX LADPROX CXMAIN OM1 LADDDIST (Features Count: 7)

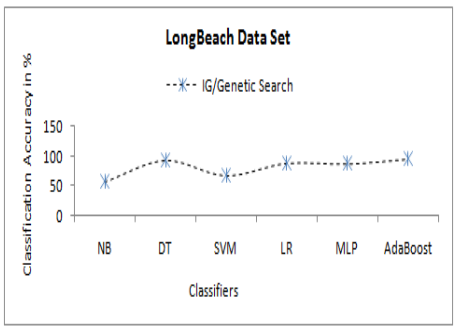
Adaboost with DT as classifier has performed best for four datasets and resulted accuracy of 98%. All the other classifiers too have shown significant improvement in the performance, with NB resulting 86% compared to the other classifiers performance of at least 90%. Despite that the Longbeach dataset and the Switzerland dataset leveraging maximum performance of 93% and 84% respectively, it is imperative that actual relevance is of 13 attributes, thus leading to study of common features across all datasets.



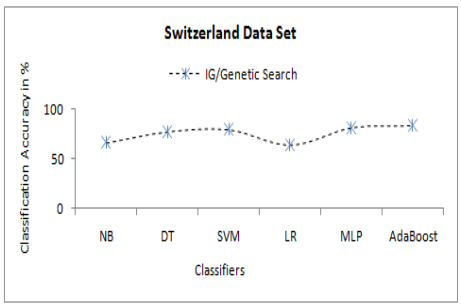
Classification Accuracy on Cleveland Dataset



Classification Accuracy on Hungarian Dataset



Classification Accuracy on Long Beach Dataset



Classification Accuracy on Switzerland Dataset

Figure 5: Classifier performance analysis on all 75 attributes

CONCLUSION

Emerging technological advancements and the kind of rapid developments that are taking place in the machine learning methods are being very resourceful to the evaluation and prediction analysis for genome structure of complex diseases. It is imperative from the review of exhaustive range of models that were proposed earlier that if predictive solutions are implemented with appropriate set of feature selection methods imbibed to the machine learning models, there are significant developments that could be envisaged in the process, and improved accuracy in terms of developments. Performance analysis that were carried out on certain feature selection models reflect that despite of the accuracy levels that are resulting from varied models, still there is scope for development and improvement of varied range of feature selection models.

REFERENCES

[1] Ashley EA, et al: Clinical assessment incorporating a personal genome. Lancet 2010, 375(9725):1525–1535.

[2] Lander ES: Initial impact of the sequencing of the human genome. Nature 2011, 470(7333):187–197.

[3] Maher B: Personal genomes: The case of the missing heritability. Nature 2008, 456(7218):18–21.

[4] Zuk O, Hechter E, Sunyaev SR, Lander ES: The mystery of missing heritability: Genetic interactions create phantom heritability. Proc Natl AcadSci U S A 2012, 109(4):1193–1198.

[5] Lehner B: Molecular mechanisms of epistasis within and between genes. Trends Genet 2011, 27(8):323–331.

[6] Moore JH, Asselbergs FW, Williams SM: Bioinformatics challenges for genome-wide association studies. Bioinformatics 2010, 26(4):445–455.

[7] Califano A, Butte AJ, Friend S, Ideker T, Schadt E: Leveraging models of cell regulation and GWAS data in integrative network-based association studies. Nat Genet 2012, 44(8):841–847.

[8] Wei Z, Wang K, Qu H-Q, Zhang H, Bradfield J, et al: From Disease Association to Risk Assessment: An Optimistic View from Genome-Wide Association Studies on Type 1 Diabetes. PLoS Genet 2009, 5(10):e1000678.

[9] 1000 Genomes Project: A map of genome variation from population-scale sequencing. Nature 2010, 467(7319):1061–1073.

[10] Mitchell, T. Machine Learning ,McGraw-Hill. This book provides a general introduction to machine learning that is suitable for undergraduate or graduate students (1997).

[11] Ohler, W., Liao, C., Niemann, H. & Rubin, G. M. Computational analysis of core promoters in the Drosophila genome. Genome Biol. 3, RESEARCH0087 (2002).

[12] Segal, E. et al. A genomic code for nucleosome positioning. Nature 44, 772–778 (2006).

[13] Bucher, P. Weight matrix description of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. J. Mol. Biol. 4, 563–578 (1990).

[14] Degroeve, S., Baets, B. D., de Peer, Y. V. &Rouzé, P. Feature subset selection for splice site prediction. Bioinformatics 18, S75–S83 (2002).

[15] Heintzman, N. et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nature Genet. 39, 311–318 (2007).

[16] Picardi, E. &Pesole, G. Computational methods for ab initio and comparative gene finding. Methods Mol. Biol. 609, 269–284 (2010).

[17] Ashburner, M. et al. Gene ontology: tool for the unification of biology. Nature Genet. 25, 25–29 (2000).

[18] Fraser, A. G. &Marcotte, E. M. A probabilistic view of gene function. Nature Genet. 36, 559–564 (2004).

[19] Beer, M. A. &Tavazoie, S. Predicting gene expression from sequence. Cell 117, 185–198 (2004).

[20] Karlic, R., R. Chung, H., Lasserre, J., Vlahovicek, K. &Vingron, M. Histone modification levels are predictive for gene expression. Proc. Natl Acad. Sci. USA 107, 2926–2931 (2010).

[21] Ouyang, Z., Zhou, Q. & Wong, H. W. ChIP–seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells. Proc. Natl Acad. Sci. USA 106, 21521–21526 (2009).

[22] Friedman, N. Inferring cellular networks using probabilistic graphical models. Science 303, 799–805 (2004).

[23] Hastie, T., Tibshirani, R. & Friedman, J. The Elements of Statistical Learning: Data Mining, Inference and Prediction (Springer, 2001). This book provides an overview of machine learning that is suitable for students with a strong background in statistics.

[24] Swan, A. L., Mobasheri, A., Allaway, D., Liddell, S. &Bacardit, J. Application of machine learning to proteomics data: classification and biomarker identification in postgenomics biology. OMICS 17, 595–610 (2013).

[25] Libbrecht, Maxwell W., and William Stafford Noble. "Machine learning applications in genetics and genomics." Nature Reviews Genetics 16.6 (2015): 321-332.

[26] Hoffman, M. M. et al. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. Nature Methods 9, 473–476 (2012).

[27] Chapelle, O., Schölkopf, B. &Zien, A. (eds) Semisupervised Learning (MIT Press, 2006).

[28] Noble, W. S. What is a support vector machine? Nature Biotech. 24, 1565–1567 (2006).

[29] Ng, A. Y. & Jordan, M. I. Advances in Neural Information Processing Systems (edsDietterich, T. et al.) (MIT Press, 2002).

[30] Wolpert, D. H. & Macready, W. G. No free lunch theorems for optimization. IEEE Trans. Evol. Comput.1, 67–82 (1997). This paper provides a mathematical proof that no single machine learning method can perform best on all possible learning problems.

[31] Urbanowicz, R. J., Granizo-Mackenzie, D. & Moore, J. H. in Proceedings of the Parallel Problem Solving From Nature 266–275 (Springer, 2012).

[32] Schölkopf, B. &Smola, A. Learning with Kernels (MIT Press, 2002).

[33] Shawe-Taylor, J. &Cristianini, N. Kernel Methods for Pattern Analysis (Cambridge Univ. Press, 2004). This textbook describes kernel methods, including a detailed mathematical treatment that is suitable for quantitatively inclined graduate students.

[34] Ramaswamy, S. et al. Multiclass cancer diagnosis using tumor gene expression signatures. Proc. Natl Acad. Sci. USA 98, 15149–15154 (2001).

[35] Urbanowicz, R. J., Granizo-Mackenzie, A. & Moore, J. H. An analysis pipeline with statistical and visualization-guided knowledge discovery for Michigan-style learning classifier systems. IEEE Comput. Intell. Mag. 7, 35–45 (2012).

[36] Tikhonov, A. N. On the stability of inverse problems. Dokl. Akad. Nauk SSSR 39, 195–198 (1943). This paper was the first to describe the now-ubiquitous method known as L2 regularization or ridge regression.

[37] Keogh, E. &Mueen, A. Encyclopedia of Machine Learning (Springer, 2011).

[38] Kruppa J, Ziegler A, König IR: Risk estimation and risk prediction using machine-learning methods. Hum Genet 2012, 131(10):1639–1654.

[39] Pahikkala T, Okser S, Airola A, Salakoski T, Aittokallio T: Wrapper-based selection of genetic features in genomewide association studies through fast matrix operations. Algorithm MolBiol 2012, 7(1):11.

[40] Okser S, Lehtimäki T, Elo LL, Mononen N, Peltonen N, et al: Genetic Variants and Their Interactions in the Prediction of Increased Pre-Clinical

Carotid Atherosclerosis: The Cardiovascular Risk in Young Finns Study. PLoS Genet 2010, 6(9):e1001146.

[41] Kooperberg C, LeBlanc M, Obenchain V: Risk prediction using genome-wide association studies. Genet Epidemiol 2010, 34(7):643–652.

[42] Clarke GM, Anderson CA, Pettersson FH, Cardon LR, Morris AP, Zondervan KT: Basic statistical analysis in genetic case-control studies. Nat Protoc 2011, 6(2):121–133.

[43] Ladouceur M, Dastani Z, Aulchenko YS, Greenwood CM, Richards JB: The empirical power of rare variant association methods: results from sanger sequencing in 1,998 individuals. PLoS Genet 2012, 8(2):e1002496.

[44] Saeys Y, Inza I, Larrañaga P: A review of feature selection techniques in bioinformatics. Bioinformatics 2007, 23(19):2507–2517.

[45] Guyon I, Elisseeff A: An introduction to variable and feature selection. J Mach Learn Res 2003, 3:1157–1182.

[46] Rakitsch B, Lippert C, Stegle O, Borgwardt K: A Lasso multi-marker mixed model for association mapping with population structure correction. Bioinformatics 2013, 29(2):206–214.

[47] Aha DW, Bankert RL: A comparative evaluation of sequential feature selection algorithms. In Learning from Data: Artificial Intelligence and Statistics V, Lecture Notes in Statistics. Edited by Fisher DH, Lenz HJ. New York: Springer-Verlag; 1996:199–206.

[48] Smialowski P, Frishman D, Kramer S: Pitfalls of supervised feature selection. Bioinformatics 2010, 26(3):440–443.

[49] Statnikov A, Wang L, Aliferis C: A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. BMC Bioinformatics 2008, 9(1):319.

[50] Castaldi PJ, Dahabreh IJ, Ioannidis JP: An empirical assessment of validation practices for molecular classifiers. Brief Bioinform 2011, 12(3):189–202.

[51] König I: Validation in genetic association studies. Brief Bioinform 2011, 12(3):253–258.

[52] Ramanan VK, Shen L, Moore JH, Saykin AJ: Pathway analysis of genomic data: concepts, methods, and prospects for future development. Trends Genet 2012, 28(7):323–332.

[53] Lee E, Chuang HY, Kim JW, Ideker T, Lee D: Inferring pathway activity toward precise disease classification. PLoSComputBiol 2008, 4(11):e1000217.

[54] Ideker T, Dutkowski J, Hood L: Boosting signal-to-noise in complex biology: prior knowledge is power. Cell 2011, 144(6):860–863.

[55] Lavi O, Dror G, Shamir R: Network-induced classification kernels for gene expression profile analysis. J ComputBiol 2012, 19(6):694–709.

[56] Mooney M, Wilmot B, The Bipolar Genome Study, McWeeney S: The GA and the GWAS: Using Genetic Algorithms to Search for Multi-locus Associations. IEEE/ACM Trans ComputBiolBioinform 2012, 9(3):899–910.

[57] Tang Chun Wong, Genetic Algorithms [Internet] 1996 [updated 1996 June 16; cited 2015 Jan 10].

[58] Seo Y-W. Class InfoGain [Internet]. 2003 [updated 2003 August 15; cited 2015 Jan 10].

[59] Harrington P. Machine Learning in Action. New York: Manning Publication, Special Sales Department; 2012.

[60] WebMD. Heart Disease Health Center [Internet] 1999 [cited 2015 January 10].