Journal of Theoretical and Applied Information Technology
31st August 2017. Vol.95. No.16

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

OPTIMIZING GENOME FEATURES USING T-TEST TO CLASSIFY THE GENE EXPRESSIONS AS CORONARY ARTERY DISEASE PRONE AND SALUBRIOUS

E. NEELIMA[1], M.S. PRASAD BABU[2]
[1]Assistant Professor, GITAM University, Department of CSE, Visakhapatnam, AP, INDIA
[2]Professor and Vice Principal, Andhra University, Department of CS&SE Visakhapatnam, AP, INDIA
E-mail: [1]eadha.neelima@gmail.com, [2]profmspbabu@gmail.com

ABSTRACT

Cardio Vascular Disease in terms of coronary artery disease and myocardial infractions are one of the majorly impacting factors towards the mortality rates. The kind of revolutionary developments that has taken place in the genomic diagnosis and the solutions that are developed for diagnosis of heart diseases based on analysis of molecular data of blood cells has improved the accuracy of diagnosis phenomenally. In recent past, analysing gene expression data and using for contemporary misnaming models. Particularly using machine learning strategies to predict and classify the given unlabelled gene expression record. In regard to this a substantial requirement is feature optimization, which is since the overall genes observed in human body are closely 25000 and among them 636 genes are cardio vascular related. Hence, it complexes the process of training the machine learning models using these entire cardio vascular gene features. Hence, this manuscript proposed the usage of ANOVA standard called t-test to select optimal features. The experimental study indicating that the number of optimal features those selected by proposed model is substantially low that compared to the other contemporary models. Divergent classifiers those trained by the features selected through proposal evinced significance in classification accuracy. We compare the results obtained from divergent classifiers those trained by the features selected using proposal and other contemporary model for performance analysis.

Keywords: *Gene Expression, Cardio Vascular Disease, Myocardial Infraction, T-Test, Coronary Artery Disease, Predictive Analysis, Genome Feature Optimization, CAD Genes, Loci, Snps*

1. INTRODUCTION

Cardiovascular diseases fall among the leading health aspects which results into so many deaths globally. Acute Myocardial infarction results from the formation of myocardial tissue as a result of reduction in the supply of blood to the heart and a number of deaths can be attributed to it. Numerous scientific studies have emphasized on solutions through diagnosis, prevention, and cure for MI.

Metabolites denote minuscule set of molecules that are resultant outcome of metabolism or the facets generated by microbes in the body [1]. Metabolites also indicate the end-products of gene expression, which is relative to the protein or certain kind of enzymatic reactions. Hence, it might be possibly leading to cellular developments in terms of pathophysiological changes.

It is imperative that metabolites that are circulated can be resourceful in predicting the presence of CAD, by identifying the disturbances that can impact atherosclerosis [2], [3].

In addition, because of the atherosclerosis resulting in blood-vessel wall interface, the measurements may plausibly impact the chronic processes in addition to predicting the existence as well the CAD stability. But for the selected metabolites that are not reviewed, incremental predictive utility like the routine clinical assessments are usually modes in addition to few researches that restricted to very few candidates based on the non-scalable techniques.

In [4]. Current high throughput and the low-cost along with high-dimensional techniques like nuclear magnetic resonance spectroscopy wherein the re-invigorated for the use of metabolic signatures towards prediction of CAD.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

ML methods are usually data-driven methods that are designed for discovering statistical patterns with high-dimensional multivariate data [5]. Supervised ML approaches signify set of techniques which are aimed at offering a function using the labelled trained set which are capable of evaluating the class input vectors.

Upon the analysis, penalized condition of logistic regression and the random forest for evaluating the predictive performance of varied metabolites of CAD for the contemporary cohort of patients referring to the hospital to investigate chest pain or the ones that are planned coronary angiography.

Annually around 17 million deaths are resulting because of CVD (Cardiovascular diseases) and it turns out to be a significant cause of mortality [6]. It is estimated that the issues like acute myocardial infraction (MI) resulting as necrosis of myocardial tissue because of reduced blood supply to the heart is a major issue in many of the heart attacks envisaged globally [6]. Despite of numerous efforts in the subject of MI related health complications, still it is turning out to be a major issue pertaining to worldwide morbidity and mortality.

Currently, various clinical symptoms are predominantly employed when diagnosing MI. some symptoms such as breathing difficulties, inconvenience or uneasiness faced by patients such as chest pain, results of abnormal Electro Cardio Gram reports, abnormal reduction in the levels of circulation of cTns (cardiac troponins) [7]. Despite the fact that there are numerous developments which have taken place within the domain, there are still some constraints and limitations which are faced when it comes to attaining accurate analysis through the use of the current diagnostic systems. For example, the contemporary techniques and solutions which were proposed in hs-cTn assays has brought about improved scope of detecting low circulating Tn concentrations. On the contrary, one of the main constraints from the process include the increase in the rates of false alarm given that higher numbers of non-diseased individuals are also indicated to be prone to conditions due to the change resulting in cTns because of other complication (this indicates reduced sensitivity) [8].

The other diagnostic measure that is adapted is the detection of cardiac mi RNAs that considered as sensitive biomarkers [9]. However, the detection accuracy is inhibited because of the poor levels of small size and tissue specific expressions. With the emergence of fast standardized and automated detection systems [10], the role of biomarkers has become very vital. BNP, CRP and other such serum inflammatory markers were also considered as a segment of cardiovascular biomarkers; however, they only could result in certain incremental improvement towards diagnosis [11], [12], [13].

Knowledge of pathological and physiological processes from established pathways are becoming significant aspect in developing many of the contemporary cardia biomarkers. But as a fact, microarray platforms in contrast consider the expression of varied genes simultaneously, ensuring gene expression profiling across various pathways in parallel. Such an approach has the scope and potential for representing a significant range of pathophysiological processes of CVD economically and more efficiently [14]. Gene expression profiling also supports in identifying potential biomarkers that were not earlier observed or associated for CVD.

In regard to this context, our earlier contribution MAGED [15] focused to devise a novel metaheuristic approach to predict Coronary artery disease, unstable angina, and myocardial infarction. The lessons learnt from the experiments and the observations of the many of existing models, it is obvious to conclude that the performance of the prediction analysis over gene profiling confines due to the complexity selection and optimizing the features from gene profiles.

2. LITERATURE REVIEW

Gene expression analysis supports in understanding and discovering contemporary and sensitive biomarkers of CVD. Many researchers have focused in the domain, and gene expression analysis has yielded close to 482 genes that are associated to the coronary atherosclerotic plaques and majority of them were not earlier linked to atherosclerosis [16]. Wide-scale gene expression profiling that was observed in 56 divergent genes for atherosclerotic and non-atherosclerotic human coronary arteries. Among them 49 were identified to be associated to CAD [17].

In [18], the researchers have focused on classifying genes, to tag age and gender correlation to obstructive CAD among the non-diabetic patients [18]. Divergent range of gene expressions were identified which discriminates ischemic and non-ischemic range of cardiomyopathy conditions

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

among the patients facing the end-stages [19], [20]. In another study of microarray analysis, gene expression profiling was adapted for discovering genes that are related to heart failure using the expression profiles of 12 patients having heart failure [21]. In [22], the study over MI patients and the normal controls depict that the genetic markets and the deregulated pathways that are associated with levels of disease recurrence in the MI patients.

Key objective of genome wide association studies (GWASs) supports in detection of traits over the population wherein variants are given genomic locations. Earlier, GWAS experiments used microarrays that were designed by many of the earlier determined variants of human population.

SNPs (Single nucleotide polymorphisms) has variations that are usually frequent in humans. Modern GWAS set of analysis adapts comprehensive set of variants and the complete set of variants along with entire set of genome sequencing data that are not confined to subset of variants.

Focus is usually on some of the key challenges of GWAS, in terms of excellent review for the present set of GWAs as depicted in [23]. In terms of data analysis, GWAS reflects upon the statistical significance amidst a casual variant in terms of risks for the individuals those are affected towards a specific conditions. However, the key issue is that in the case of GWAS and the association based solutions like eOTLs, the fact is that it points to correlation and not to the cause.

Due to the confounding set of hidden variables and the correlations amidst nearby variants that are brought by cross-over or in terms of differences in terms of subpopulations. [24].

In the GWAS study, even the casual variant is not observed. Also, GWAS offers high quantum of putative casual mutations. As the researchers, might be focusing on the candidates in terms of issues of "narrative potential" [25].

It is very essential that some of the GWAS research targets tens of millions of SNPs, in terms of thousands of individuals. Evaluation of statistical importance in terms of number of SNPs that are highly challenging, which usually needs significant multiple-hypothesis correction and false discovery rate analysis [26].

Problem is usually compounded in terms of common variants envisaging weaker impacts and the ones that comprise strong impacts, as they tend to be very rate and pointed [27]. For rising the significance, many studies have targeted SNPs of genome's coding region [28] that has high probability of altering the function of proteins as depicted in [29].

Other intrinsic way of addressing the challenge is focusing on the size of sample. Significant quantum of resources are generally channelled for big population study which focus on TCGA, Hap Map which are generally lead to debates among the research community that concerns the cost-benefit ratio for the projects [30] [31] [32].

Other key issue is the structure of population and the stratification envisaged in the process. In [33], a genetic classifier depending on SNPs for detection of ASD were targeted that leads to complexities [34]. It is also envisaged that many signals that were observed due to aspects like potential population stratification.

Usage of GWAS data beyond the statistical associations are usually based on computational models that consider the SNP profiles of individuals as input, to predict risks of diseases.

SNP profiles usually tend to have higher dimensional sets which big proportions of SNPs that are not relevant to diseases that are not considered.

Various tools that are available for prioritizing usual variants which were proposed in [35], SIFT [36], SPANR [37]. It is also essential to focus on machine learning models that are usually employed for learning the predictive models of disease risks as discussed in [38] and [39].

The question of how the blood transcriptase could denote the transcriptional changes in the heart has been focused in [40], wherein a genome wide survey using microarrays were calculated and sequence tags were expressed and the blood transcript was compared to the transcript of human tissues including the heart. Results of the study depict that close to 80% of the given tissue and 84 of it with heart, reflects peripheral blood transcriptase. It can be an economic and easily accessible tool for proxy gene expression in other tissues [40].

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

In [41], the research has detected different kinds of issues of imbalances which may creep up in the use of microarray as a result of noisy, huge volume, as well as irrelevant samples. Due to the afore-stated complexities, researchers have always embarked on the application of swarm intelligence tools so as to address the given issue. The research used the ACO sampling technique that developed based on Ant Colony Optimization (ACO) algorithm for the elimination of the noisy, as well as the irrelevant features during the feature selection process.

SVM classifiers were used due to its prominence for high dimensional classification of data even with small sample set. Concerns regarding unstable classification performance noted in cross validation process forms part of the main factors. In [42], the authors have embraced a hybrid model for choosing optimal features through the use of Artificial Bee Colony (ABC) and the classification was conducted through the use of SVM classifiers. ABC was used for clustering, as well as for the selection of optimal features that minimizes search space. Various experimental studies are depicting unstable accuracy at the 10-fold classification level.

The model in [43], is proposing the application of ACO, as well as RST (Roughest Theory) to attain the optimized feature count. The accuracy of feature selection is inversely proportional to the levels of dimensionality in the feature set. At the same time, in [44], it is exploring the application of BAT algorithm for the reduction of the dimensionality of feature, as well as the selection of optimal features.

At the same time, in [45], fuzzy based model which depicts rules based on the relationship among the developed features through the use of a mix of BAT and ACO technique. Additionally, the rules may be in use for the selection of optimal features in a manner that is highly dynamic. Among the constraints which envisaged in the given model, exposure is required so as to make sure that there is selection of prior attributes which helps in the selection of dependent attributes based on devised fuzzy rules.

BCO and RST combined in [46] in which there is much focus on clustering features on the basis of the phenotype or based on the pattern which points out the optimal features. It employed LSDA (Locality Sensitive Discriminant Analysis) for minimizing dimensionality of feature sets, which further clusters, the outcome through the use of FCM algorithm. FCM was used in together with the ABC approach for feature similarity evaluation during the formation of the clusters.

FCM incorporated with Artificial Bee Colony (ABC) approach which is employed for feature similarity evaluation in the process of the formation of the cluster. Other contemporary models in feature optimization include in [47] Binary Bat Algorithm (BBA), as well as ABC were employed and in [48] m Minimum Redundancy and Maximum Relevance (RMR), and PSO, as well as DT in [49].

Though varies studies have focused on gaining insights in to differential expression in cardiovascular outcomes, every attempt is to focus on information to classify the patient records to an outcome [50]. Such an approach leads to offer diagnostic tool for sub-classification of patients.

In [50], the discriminatory features for differentiating among the patients prone to MI and the normal people, people identified for CAD and the unstable angina were explored using the gene expression in the blood cells. Blood transcriptase that is adapted easily access the tissue for diagnostic purposes. Majority of them reflect the issues of computational complexity resulting from dense number of gene features that are used in the learning process.

It is essential to focus on comparative genomics that are high means for tracing genomic sequences with function. Sequence conservation is one such solution. Slow accumulation of random mutations and the selective pressures against mutations which might be damaging for reproductive fitness in the population [51].

Genomes of varied species that are diverging, in terms of random mutations that consumes more time. In comparison of genomes, varied long distinct sequences that are identical and the ones that are conserved. Upon confirming a sequence, there is strong evident of selective pressure over the positions that are in sequences. Many studies emphasize that around 5 to 6% of human genome is considered for mammals [52] [53].

Detection of conserved sequences usually have highly instrumental in annotation of functional components in human genome [52], like exons. Scores of conservation are usually available for numerous organisms from software too phast [54].

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

It is essential to pointing out that for every position of genome, phast Cons provides a number between 0 and 1, wherein 0 denotes hardly any discernible conservation, whereas 1 points 100 percent conservation to every specie that takes in to account.

Other techniques for quantification conservation usually comprise GERP [55] and phylo P [56]. Every conservation score for position of human genome might be considered as "track" in the UCSC Genome Browser [57].

Mutation that lowers reproductive fitness is seen as deleterious, wherein a mutation causing a disease is pointed to pathogenic [25]. Numerous mutations are both deleterious and pathogenic. It is essential to comprehend wherein conservation only provides information pertaining to deleteriousness.

Also, in the other dimension, conservation-based methods usually have high impact feature for predictive models. An exemplary model is about combining annotation dependent depletion [58].

Mutation Simulator results in generation of realistic synthetic mutations that do not comprise selective pressure.

Machine learning and deep learning has to attain human-level performance in the case of natural language processing and speech recognition too. In parlance, even in the case of genome biology too, it is distinct to the domains in significant manner. Humans perform effectively in terms of human action.

Despite that there is hardly any evolutionary pressure for humans in terms of interpretation, perception and towards responding to patterns of light that is produced, wherein humans has the ability for interpreting the genome.

As a consequence, it is very important to focus on the contemporary set of biological knowledge and also the data in to learning. Despite that the possibility of association of genome to certain diseases might be very complex, still it can't be ruled out in terms of contrast to speech or image recognition, wherein the prediction should be provided with input.

Possibility in terms of association of genome to few diseases might be complex as it can't be modelled considering the set of practical quantity of inputs. Such a sharp contrast to speech or image recognition wherein the prediction is needed as input. Additionally it is essential to point that due to inherent stochastic city of cellular processes, environmental factors vary from one level to other. [59].

In terms of machine learning, it is essential to focus on few similarities amidst genome biology and the other domains. Such relationship might be considered as landscape comprising genotype results wherein distinct kinds of phenotypes are pointed [60].

In [61] and [62] similar observations were made for varied applications the other application domains, within the semantics change significantly. In terms of focusing on deep learning, consistent outcome in terms of vision and speech is considered.

Among the exciting elements, current succeeds in terms of 'end-to-end' learning wherein various layers has the efficacy of learning from extremely low level [63] [64]. In comparison of progress, in terms of genomics, potential effects of genomic medicine is pragmatic.

Variables are hard to evaluate in comparison to the phenotypic observations. There is need to evaluate measuring certain variables in every patient for a big set of population. It is considered that latter approach provides better sense of deciphering genomic instructions of cell, wherein more information is available pertaining to biological mechanisms which play a vital role.

"Genomic invariance" assumption is carried out that implies assumption that regulatory processes act similar across whole genome. We can DNA to cell variable relationship by diverse range of genome for independent metrics.

Huge growth in terms of genomic data usually considers privacy, storage and computing challenges that make it more complex for small research groups to work on large scale.

In computation, parallel-distributed training solution that addresses big data sets were pointed out [65] [66]. Algorithms having effectiveness in terms of computing clusters with 1000 nodes. In a recent illustration having large-scale machine learning might be very steady and in a manner comprising economical over consumer off-the-shelf clusters accelerated by GPUs [64] and [67].

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

The process of cloud computing has facilitating reproducible research using the virtual machine images. In [68], WSSE technique wherein the data sets, pipelines and code libraries along with experimental results are packaged and accessed for inspection and follow-up research.

Reduction of sequencing costs that outpaces information technology as storage support in various conditions, wherein the conventional use case wherein practitioners focus on genomic data for local computer. [69] and [70].

Configuration of large-scale machine learning solutions appropriate level of engineering expertise. Pipelines resulting in brittle complex an out dated dependencies with production is highly taxing [71].

In addition to genomic and disease risk data can be very resourceful for leveraging cell-level data and explicit model of manner, wherein genome affect them.

Main intension of measuring variables for measuring technology and thereafter predicting variables might be very expensive [72]. Novel mRNA isoforms, by long-read technology [73], [74], [75], as well as mRNA levels in the single cells [76], [77], [78].

Integration of varied emerging data sources in to analysis turns out to be a key challenge. Emergence of gold standard [79] denotes that despite of hurdles, in comparison to possible value of transforming medicine for saving lives.

Whereas ML techniques predicted the presence or absence of CAD in unadjusted approaches (through the use of metabolite data only) with very high accuracy or sensitivity, in adjusting for confounders they were generally outperformed by the PCA regression in terms of ROC AUC as well as accuracy. This generally suggests that a small quantity of metabolites is capable of being included in the prediction models. Several individual metabolites which were established to be statistically important are in line with the past literature and our pathological comprehension of CAD as well as its development. The hero genic lipid particles like LDL are generally known to be causally linked to atherosclerosis, whereas others like creatinine generally reflect renal function. They are also established markers of CAD risk. Numerous other metabolites are having no past robust association with CAD which includes lactate and phenylalanine. They generally represent potentially new investigation avenues. In seeking metabolic signature so as to predict CAD, the ML models generally suffered from low specificity. The exploratory analysis has pointed out besides exemplifying the value of the ML models for CAD prediction though the use of high-dimensional data. It has illustrated that that accuracy of the conventionally regression-based approaches may be surpassed. More research is, however, needed before the methods can be translated into clinical solutions.

In this paper, the proposed solution is a devised feature optimization strategy which is based on t-test, which can significantly reduce feature counts when compared to many of the existing solutions. Despite of reducing the feature counts used for analysis, still the levels of accuracy in classification has increased significantly and the false alarm rates has dwindled to desired levels.

3. GENOME FEATURE OPTIMIZATION CLASSIFYING GENE EXPRESSIONS AS CORONARY ARTERY DISEASE PRONE AND SALUBRIOUS

In this section of the study, the process of optimizing the genome features and classification of gene expressions for the coronary artery disease prone and salubrious. Initially, the emphasis is on exploring the methods and materials used for devising the model. Furtherly, the method of feature optimization based on ANOVA standard t-test is considered, which is followed by the exploration of process.

3.1 Methods and Materials

In this section, the methods and the materials for the proposed model is discussed. The sub-section details the feature set, process of t-test for selecting optimal features and the classifiers used for the process is discussed.

3.1.1 The feature set

636 of the total 25000 genomes available, are of cardio vascular related genomes [80] that are termed as CAD genes. The process of complexity for estimating the correlation amidst 636 genome features are considerably high and it could lead to more false alarm rates in the prediction models. Hence to reduce the complexities of feature dimensionality and ensure process complexity is low and linear, the false alarming rates have to be low to great extent possible.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Every record of the dataset used for training and testing phases comprise SNP (single nucleotide polymorphism) for every gene. It denotes that the genetic variation among respective genes. Initial length of every record is 636 values denoting the SNPs of all the 636 genomes that are listed in CAD Genes.

The initial dataset comprises set of records labelled as salubrious (ones that do not have any evidence of coronary artery diseases) and the other set labelled as prone (prone to coronary artery diseases).

Dimensionality of the genes count should be limited from 636 to the minimal levels. Hence, ANOVA standard t-test is used in the process of reducing dimensionality and to optimize gene count for building the proposed scale. In the following section, the inputs on the t-test adapted for improving the feature optimization is discussed.

3.1.2 t-Test

The adaptation of this method is observed as significant since, the given gene profiles for training are representing same distribution and contemporary statistics indicating that t-test is optimal to identify the two different sets of values from the same distribution are distinct or similar [81]. Attributes in every record of a chosen dataset indicate that every gene of CAD Genes with count of 636. Hence, every record comprises 636 SNPs as values pertaining to all genomes.

For defusing the genes count and to use optimal features, the covariance amidst the values denoting every gene in the records are labelled as prone and salubrious for every feature. Genes are optimal features comprising significant covariance amidst the values pertaining to prone and salubrious records.

For estimating the variance of SNPs which exists as values of gene as prone and salubrious records for a selected training set, ANOVA standard t-test is used for the process. In [82], [83], the method is used for analysis which advocates the resourcefulness of the ANOVA t-test. The t-score is adapted for selecting optimal features pertaining to disease prone and salubrious records of the training set.

Diversity of values in to two varied sectors denotes by t-score is denoted in the following equation

$$t-score=\frac{(M_{v1}-M_{v2})}{\sqrt{\frac{\sum_{i=1}^{|v1|}(x_i-M_{v1})^2}{|v1|-1}+\frac{\sum_{j=1}^{|v2|}(x_i-M_{v2})^2}{|v2|-1}}}$$

In the equation above

- M_{v1}, M_{v2} Denotes the mean of the values perceived in corresponding vectors $v1, v2$ and these vectors denotes the SNPs that exists as values for a gene in corresponding records that are labelled as prone and salubrious accordingly in the training set.
- The notations x_i, x_j denotes each SNP of respective vectors $v1, v2$ of corresponding sizes $|v1|, |v2|$

T-score is the ration amidst mean differences of corresponding vectors and the square root of the cumulative of mean square distances of respective vectors.

Further, the degree probability (p-values) [84] for t-table [85] is obtained. The p-value with lesser probability threshold depicts that both the vectors are very distinct and feature representing the corresponding vectors are optimal featured.

3.1.3 Classifiers

The classifiers that are vividly used in the contemporary research are chosen for classifying the selected unlabelled microscopic images in order to ensure that malaria infected and the non-malaria infected are identified appropriately. SVM [86], Naïve Bayes [87] and Adaboost [88] were adapted, which are adapted by the most of contemporary classification models in recent literature [89].

While SVMs (Support Vector Machines) are near optimal for classification process. The other popular classifier used in the process is the NB (Naïve Bayes). NB is much easier for building and can be very effective if right kind of inputs are provided for features to the independent classes. Adaboost is the other classifier which is profoundly used which boosts the performance of decision trees, and hence the miss classification rate in the process is very low. In a comparative analysis of all the three classifiers (SVM, NB and AdaBoost) the miss-calculation rates are very low with SVM and NB classifiers. Process completion time observed for SVM, NB were predominantly

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

stable and low when compared to the AdaBoost classifier's performance.

3.1.4 The dataset

The Dataset is prepared from the records representing the coronary artery susceptibility mode (NCBI GEO Dataset ID: GDS4527) and Atherosclerotic Coronary Artery Disease prone (NCBI GEO Dataset ID: GDS3690) that are collected from NCBI gene expression omnibus (NCBI GEO) [90], which is gene expression dataset repository. The dataset GDS4527 contains gene expressions of 20 subjects. Among these 10 records labelled as salubrious and rest of the records labelled as prone to coronary artery disease. The other dataset GDS3690 contains 153 records and among these, 66 records labelled as salubrious, and rest 87 records labelled as prone to coronary artery disease. From the records of these two datasets that are together representing 173 subjects, the values observed for CAD Genes (total 636 genes) collected as a record for each subject. The statistics of the final dataset that prepared depicted in Table 1.

Table 1: Dataset Statistics

Total number of records	173
Each record length	636
Number of records with salubrious label	76
Number of records with disease prone label	97

3.2 Optimizing Genome Features

Initially partitions the overall labelled records in the given dataset G as two sets P, S representing coronary artery disease prone and salubrious records respectively. The set P and set S are in the form of matrix of size records count as row count and CAD genes count as column count that fixed to 636 [80]. Each row of the matrix is a vector that represents the SNPs obtained for all CAD Genes in respective to an individual case and each column is a vector that represents the SNPs obtained for a specific gene in all given cases. The context of the optimal feature selection is that a gene having diversified vector of SNPs in respective of prone and salubrious record sets P, S. Further applies t-test on the value obtained for a gene in regard to prone and salubrious sets as follows

step 1: $\forall_{i=1}^{|C|}\{pc_i \exists pc_i \in P \wedge sc_i \exists sc_i \in S\}$ Begin

step 2: $\langle pc_i \rangle = \frac{\sum_{k=1}^{|pc_i|}\{pc_i(k)\}}{|pc_i|}$ // finding the mean $\langle pc_i \rangle$ of the all values exist in column vector pc_i of the set P that represents SNPs found in all records of the set P for gene pc_i

step 3: $\langle sc_i \rangle = \frac{\sum_{k=1}^{|sc_i|}\{sc_i(k)\}}{|sc_i|}$ // finding the mean $\langle sc_i \rangle$ of the all values exist in column vector sc_i of the set S that represents SNPs found in all records of the set S for gene sc_i

step 4: $rmsd_{pc_i \square sc_i} = \sqrt{\frac{\sum_{k=1}^{|pc_i|} pc_i(k) - \langle pc_i \rangle}{|pc_i| - 1} + \frac{\sum_{k=1}^{|sc_i|} sc_i(k) - \langle sc_i \rangle}{|sc_i| - 1}}$ // finding the root mean square distance $rmsd_{pc_i \square sc_i}$ of the vectors pc_i and sc_i

step 5: $t_{pc_i \square sc_i} = \frac{(\langle pc_i \rangle - \langle pc_i \rangle)}{rmsd_{pc_i \square sc_i}}$ // Assessing the t-score of the vector pc_i and vector sc_i comparison

step 6: $if\left(p\left(t_{pc_i \square sc_i}\right) < pt\right)$ // If the degree of probability $p\left(t_{pc_i \square sc_i}\right)$ observed for $t_{pc_i \square sc_j}$ is less than the probability threshold (usually 0.01, 0.05 or 0.1) given

step 7: $oGene \leftarrow C\{i\}$ // then the i^{th} gene of the CAD Genes set C considers as optimal and moved to the optimal gene set $oGene$

step 8: End

4. EXPERIMENTAL STUDY

The experiments conducted to assess the classification accuracy and computational efficiency of the proposed feature optimization strategy. In related to this, the experiments conducted on dataset that explored in earlier section (see sec 3.1.4). The classifiers that used to classify the unlabelled records using the optimal gene features selected by proposed model are naïve bays, SVM and Adaboost. The classification assessment standards [91] like precision, sensitivity, specificity, and accuracy used to assess the significance of the optimal feature selection model proposed towards predicting the record status as prone to coronary artery disease or

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

salubrious. In regard to estimate the significance of the proposed optimal feature selection model, the classification results were compared to the results obtained from the feature optimization technique, which is the combination of P-Value based Selection and LOOCV t-tests (leave one out cross validation t-test) that explored in [50].

4.1 Performance Analysis

The number of optimal features selected by proposed model is 18, whereas 25 is the count of optimal features selected by contemporary model [50]. The classification accuracy estimated through 4-fold classification using Naive bays, SVM and Adaboost classifiers. The table 2, table 3, table 4, and table 5 depicts performance statistics of respective folds.

Table 2: Performance Statistics Obtained from Three Classifiers Applied On Features Selected By Proposed (T-Test) and Contemporary Model 18 from Fold 1 of Dataset

	Naïve Bays		SVM		Adaboost	
	t-test	p-value [50]	t-test	p-value [50]	t-test	p-value [50]
TP	22	18	23	19	23	20
FP	1	5	2	5	1	3
TN	18	14	17	14	18	16
FN	3	7	2	6	2	5
precision	0.957	0.783	0.92	0.792	0.958	0.87
Sensitivity	0.88	0.72	0.92	0.76	0.92	0.8
Specificity	0.947	0.737	0.895	0.737	0.947	0.842
Accuracy	0.909	0.727	0.909	0.75	0.932	0.818

Table 3: Performance Statistics Obtained from Three Classifiers Applied on Features Selected by Proposed (T-Test) and Contemporary Model 18 From Fold 2 of Dataset

	Naïve Bays		SVM		Adaboost	
	t-test	p-value	t-test	p-value	t-test	p-value
TP	23	18	23	18	24	20
FP	2	6	1	4	1	2
TN	17	13	18	15	18	17
FN	2	7	2	7	1	5
precision	0.92	0.75	0.958	0.818	0.96	0.909
Sensitivity	0.92	0.72	0.92	0.72	0.96	0.8
Specificity	0.895	0.684	0.947	0.789	0.947	0.895
Accuracy	0.909	0.705	0.932	0.75	0.955	0.841

Table 4: Performance Statistics Obtained from Three Classifiers Applied on Features Selected by Proposed (T-Test) and Contemporary Model [50] from Fold 3 of Dataset

	Naïve Bays		SVM		Adaboost	
	t-test	p-value	t-test	p-value	t-test	p-value
TP	24	23	23	20	23	21
FP	3	5	2	4	1	3
TN	16	14	17	15	18	16
FN	1	2	2	5	2	4
precision	0.889	0.821	0.92	0.833	0.958	0.875
Sensitivity	0.96	0.92	0.92	0.8	0.92	0.84
Specificity	0.842	0.737	0.895	0.789	0.947	0.842
Accuracy	0.909	0.841	0.909	0.795	0.932	0.841

Table 5: Performance Statistics Obtained from Three Classifiers Applied on Features Selected By Proposed (T-Test) and Contemporary Model [50] From Fold 4 of Dataset

	Naïve Bays		SVM		Adaboost	
	t-test	p-value	t-test	p-value	t-test	p-value
TP	23	20	23	22	23	22
FP	2	6	1	5	1	3
TN	17	13	18	14	18	16
FN	2	5	2	3	2	3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

precision	0.92	0.769		0.958	0.815		0.958	0.88
Sensitivity	0.92	0.8		0.92	0.88		0.92	0.88
Specificity	0.895	0.684		0.947	0.737		0.947	0.842
Accuracy	0.909	0.75		0.932	0.818		0.932	0.864

The classification accuracy observed for given 4-fold data is consistent for the optimal gene features selected by proposed model, whereas considerable inconstancy noticed in classification accuracy from the features selected by contemporary model [50], which is depicted in figure 1. The process complexity of the respective classifiers is much lower while classifying the test data using the features selected by proposed model that compared to the process time taken by classifiers to classify given data using features selected by contemporary model [50]. This is since the number of features selected from t-test is 18 that significantly lower than the number features selected by contemporary model, which are 25 in count. The figure 2 depicts the comparison of process completion time observed for divergent classifiers with optimal features selected from both t-test and contemporary model p-value.

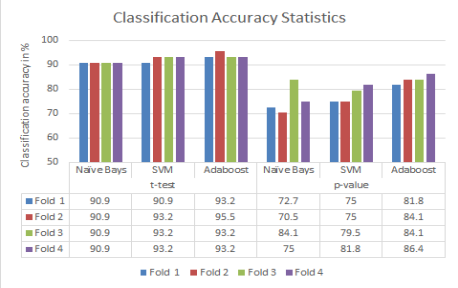

Figure 1: Classifier Accuracy Statistics Observed From Classifiers Applied on Features Selected By Proposed (T-Test) and Contemporary Model (P-Value) [50]

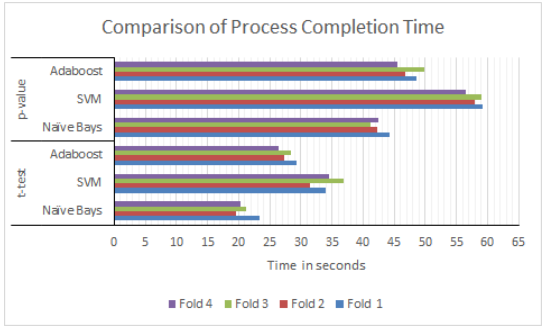

Figure 2: Process Completion Time Statistics Observed from Classifiers Applied on Features Selected By Proposed (T-Test) and Contemporary Model (P-Value) [50]

Among the three classifiers, the Adaboost outperformed (with an average of classification accuracy 94% for t-test features and 84% for p-value features) the other two classifiers Naïve bays (with average of 91% classification accuracy for t-test features and 76% for p-value features) and SVM (with average of 92% classification accuracy for t-test features and 78% for p-value features). The overall classification performance of the naïve bays is low that compare to other two.

5. CONCLUSION

This manuscript depicted a statistical approach to discover optimal features from gene expressions those formed by cardio vascular genomes listed in CADGenes records. The model proposed is built using ANOVA standard t-test [50], which is a standard scale to identify the distance between two vectors having pooled values. The process of optimal features selection among the 636 Cardio vascular related genomes is,

i. The proposed model pools the values observed for a specific genome from the records labelled as disease prone as a vector,
ii. Also pools the values observed for same genome in set of records labelled as salubrious and estimates the t-score between these two vectors using the t-test.
iii. If t-score is significant according to degree of probability (p-value), then that genome selects as optimal

The experimental study evincing that the proposed model is significant as it selects minimal number of features (18 genomes) as optimal and at improved the classification accuracy of the

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

different classifiers. The performance of the proposed model scaled by comparing with results obtained from other contemporary model that selected 25 features as optimal at failed to maintain the consistency in classification accuracy of the divergent classifiers. In order to devise the computer aided prediction system, future research would contribute a swarm intelligence approach such as cuckoo search as a Boolean classifier to predict the given record is prone to disease or salubrious.

REFERENCES:

[1] Villas-Bôas, Silas G., et al. "Microbial Metabolomics: Rapid Sampling Techniques to Investigate Intracellular Metabolite Dynamics—An Overview." Metabolome Analysis: An Introduction (2006): 203-214.

[2] Würtz P, Mäkinen VP, Soininen P, et al. Metabolic Signatures of Insulin Resistance in 7,098 Young Adults, Diabetes 61 (2012), 1372-1380.

[3] Wang T, Larson M, et al. Metabolite profiles and the risk of developing diabetes, Nat Med 17 (2011), 448– 453.

[4] Soininen P, Kangas A, Würtz P, et al., Quantitative Serum Nuclear Magnetic Resonance Metabolomics in Cardiovascular Epidemiology and Genetics, Circulation: Cardiovascular Genetics 8 (2015), 192-206.

[5] Goldstein B, Navar AM, Carter R, Moving beyond regression techniques in cardiovascular risk prediction, EHJ (2016), 10.1093/eurheartj/ehw302.

[6] Mozaffarian D, Benjamin EJ, Go AS, Arnett DK, Blaha MJ, Cushman M, et al. Heart disease and stroke statistics—2015 update: a report from the American Heart Association. Circ. 2015; 131:29–32.

[7] Thygesen K, Alpert JS, White HD, Jaffe AS, Apple FS, Galvani M, Katus HA, Newby LK, Ravkilde J, Chaitman B, Clemens PM. Universal definition of myocardial infarction Kristian Thygesen, Joseph S. Alpert and Harvey D. White on behalf of the Joint ESC/ACCF/AHA/WHF Task Force for the Redefinition of Myocardial Infarction. European heart journal. 2007 Oct 1; 28(20):2525-38.

[8] Eggers KM, Lind L, Venge P, Lindahl B. Will the universal definition of myocardial infarction criteria result in an over diagnosis of myocardial infarction? The American journal of cardiology. 2009 Mar 1; 103(5):588-91.

[9] Wang Z, Luo X, Lu Y, Yang B. miRNAs at the heart of the matter. J of Mol Med. 2008; 86:771–783.

[10] de Planell-Saguer M, Rodicio MC. Detection methods for microRNAs in clinic practice. Clin Biochem. 2013; 46:869–878. doi: 10.1016/j.clinbiochem.2013.02.017 PMID: 23499588

[11] Melander O, Newton-Cheh C, Almgren P, Hedblad B, Berglund G, Engström G, et al. Novel and conventional biomarkers for prediction of incident cardiovascular events in the community. The J of the Amer Med Assoc. 2009; 302:49–57.

[12] Shah T, Casas JP, Cooper JA, Tzoulaki I, Sofat R, McCormack V, et al. Critical appraisal of CRP measurement for the prediction of coronary heart disease events: new data and systematic review of 31 prospective cohorts. Inter J of Epid. 2009; 38:217–231.

[13] Wilson PWF, Pencina M, Jacques P, Selhub J, D'Agostino R, O'Donnell CJ. C-reactive protein and reclassification of cardiovascular risk in the Framingham Heart Study. Circ: Card Qual and Outc. 2008; 2:92–97.

[14] Pedrotty DM, Morley MP, Cappola TP. Transcriptomic biomarkers of cardiovascular disease. Prog in Card Dis. 2012; 55:64–69.

[15] Neelima, E., and MS Prasad Babu. "MAGED: Metaheuristic Approach on Gene Expression Data: Predicting the Coronary Artery Disease and the Scope of Unstable Angina and Myocardial Infarction." Global Journal of Computer Science and Technology 16.4 (2016).

[16] Randi AM, Biguzzi E, Falciani F, Merlini P, Blakemore S, Bramucci E, et al. Identification of differentially expressed genes in coronary atherosclerotic plaques from patients with stable or unstable angina by cDNA array analysis. J of Throm and Haem. 2003; 1:829–835.

[17] Archacki S, Angheloiu G, Tian XL, Tan FL, DiPaola N, Shen GQ, et al. Identification of new genes differentially expressed in coronary artery disease by expression profiling. Phys Genom. 2003; 15:65–74.

[18] Elashoff MR, Wingrove JA, Beineke P, Daniels SE, Tingley WG, Rosenberg S, et al. Development of a blood-based gene expression algorithm for assessment of obstructive coronary artery disease in nondiabetic patients. BMC Med Genom. 2011; 4:4–26.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[19] Kittleson MM, Ye SQ, Irizarry RA, Minhas KM, Edness G, Conte JV, et al. Identification of a gene expression profile that differentiates between ischemic and nonischemic cardiomyopathy. Circ. 2004; 110:3444–3451

[20] Kittleson MM, Minhas KM, Irizarry RA, Ye SQ, Edness G, Breton E, et al. Gene expression analysis of ischemic and nonischemic cardiomyopathy: shared and distinct genes in the development of heart failure. Phys Genom. 2005; 21:299–307.

[21] Min KD, Asakura M, Liao Y, Nakamaru K, Okazaki H, Takahashi T, et al. Identification of genes related to heart failure using global gene expression profiling of human failing myocardium. Bioch and Biophy Res Comm. 2010; 393:55–60.

[22] Suresh R, Li X, Chiriac A, Goel K, Terzic A, Perez-Terzic C, et al. Transcript me from circulating cells suggests dysregulated pathways associated with long-term recurrent events following first-time myocardial infarction. J of Mol and Cell Card. 2014; 74:13–21.

[23] P. M. Visscher, M. A. Brown, M. I. McCarthy, and J. Yang, ‘‘Five years of GWAS discovery,’’ Amer. J. Human Genetics, vol. 90, no. 1, pp. 7–24, 2012.

[24] J. K. Pritchard and M. Przeworski, ‘‘Linkage disequilibrium in humans: Models and data,’’ Amer. J. Human Genetics, vol. 69, pp. 1–14, 2001.

[25] D. G. MacArthur et al., ‘‘Guidelines for investigating causality of sequence variants in human disease,’’ Nature, vol. 508, no. 7497, pp. 469–476, 2014.

[26] R. C. Johnson et al., ‘‘Accounting for multiple comparisons in a genome-wide association study (GWAS),’’ BMC Genomics, vol. 11, p. 724, 2010.

[27] G. Gibson, ‘‘Rare and common variants: Twenty arguments,’’ Nature Rev. Genetics, vol. 13, no. 2, pp. 135–145, 2012.

[28] Bamshad, Michael J., et al. "Exome sequencing as a tool for Mendelian disease gene discovery." Nature reviews. Genetics 12.11 (2011): 745.

[29] Ng, Sarah B., et al. "Targeted capture and massively parallel sequencing of twelve human exomes." Nature 461.7261 (2009): 272.

[30] Ledford, Heidi. "End of cancer atlas prompts rethink: geneticists debate whether focus should shift from sequencing genomes to analysing function." Nature 517.7533 (2015): 128-130.

[31] International HapMap Consortium. "A haplotype map of the human genome." Nature 437.7063 (2005): 1299.

[32] International HapMap 3 Consortium. "Integrating common and rare genetic variation in diverse human populations." Nature 467.7311 (2010): 52.

[33] Skafidas, Efstratios, et al. "Predicting the diagnosis of autism spectrum disorder using gene pathway analysis." Molecular psychiatry 19.4 (2014): 504.

[34] Robinson, E. B., et al. "Response to ‘Predicting the diagnosis of autism spectrum disorder using gene pathway analysis'." Molecular psychiatry 19.8 (2014): 860.

[35] Adzhubei, Ivan A., et al. "A method and server for predicting damaging missense mutations." Nature methods 7.4 (2010): 248-249.

[36] Kumar, Prateek, Steven Henikoff, and Pauline C. Ng. "Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm." Nature protocols 4.7 (2009): 1073-1081.

[37] H. Y. Xiong et al., ‘‘The human splicing code reveals new insights into the genetic determinants of disease,’’ Science, 2014 vol. 347, no. 6218, DOI: 10.1126/science.1254806.

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

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

[40] Liew CC, Ma J, Tang HC, Zheng R, Dempsey AA. The peripheral blood transcript me dynamically reflects system wide biology: a potential diagnostic tool. The J. of Lab and Clin Med. 2006; 147:126– 132.

[41] Yu H, Ni J, Zhao J. ACO Sampling: An ant colony optimization-based under sampling method for classifying imbalanced DNA microarray data. Neuro computing. 2013 Feb 4; 101:309 -18.

[42] Uzer MS, Yilmaz N, Inan O. Feature selection method based on artificial bee colony algorithm and support vector machines for medical datasets classification. The Scientific World Journal. 2013 Jul 28; 2013.

[43] Arafat H, Elawady RM, Barakat S, Elrashidy NM. Using rough set and ant colony optimization in feature selection. International Journal of Emerging Trends & Technology in Computer Science (IJETTCS). 2013 Jan; 2 (1).

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[44] Taha AM, Tang AY. Bat algorithm for rough set attribute reduction. Journal of Theoretical and Applied Information Technology. 2013 May 10; 51(1):1-8.

[45] Kumar PG, Vijay SA, Devaraj D. A hybrid colony fuzzy system for analyzing diabetes microarray data. In Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2013 IEEE Symposium on 2013 Apr 16 (pp. 104-111). IEEE.

[46] Sathishkumar K, Thiagarasu V, Ramalingam M. An efficient artificial bee colony and fuzzy c means based clustering gene expression data. International Journal of Innovative Research in Computer and Communication Engineering. 2013 Jul; 1(5).

[47] Nakamura RY, Pereira LA, Rodrigues D, Costa KA, Papa JP, Yang XS. Swarm Intelligence and Bio-Inspired Computation: 9. Binary Bat Algorithm for Feature Selection. Elsevier Inc. Chapters; 2013 May 16.

[48] Alshamlan H, Badr G, Alohali Y. MRMR-ABC: a hybrid gene selection algorithm for cancer classification using microarray gene expression profiling. BioMed research international. 2015 Apr 15; 2015.

[49] Chen KH, Wang KJ, Tsai ML, Wang KM, Adrian AM, Cheng WC, Yang TS, Teng NC, Tan KP, Chang KS. Gene selection for cancer identification: a decision tree model empowered by particle swarm optimization algorithm. BMC bioinformatics. 2014 Feb 20; 15(1):49.

[50] Kazmi, Nabila, and Tom R. Gaunt. "Diagnosis of coronary heart diseases using gene expression profiling; stable coronary artery disease, cardiac ischemia with and without myocardial necrosis." PloS one 11.3 (2016): e0149475.

[51] A. Ureta-Vidal, L. Ettwiller, and E. Birney, ‘‘Comparative genomics: Genome-wide analysis in metazoan eukaryotes,’’ Nature Rev. Genetics, vol. 4, no. 4, 2003 pp. 251–262,.

[52] K. Lindblad-Toh et al., ‘‘A high-resolution map of human evolutionary constraint using 29 mammals,’’ Nature, vol. 478, no. 7370, 2011: pp. 476–482.

[53] E. T. Dermitzakis et al., ‘‘Evolutionary discrimination of mammalian conserved non-genic sequences (CNGs),’’ Science, vol. 302, no. 5647, 2003: pp. 1033–1035.

[54] A. Siepel et al., ‘‘Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes,’’ Genome Res., vol. 15, no. 8, (2005): pp. 1034–1050.

[55] Cooper, Gregory M., et al. "Distribution and intensity of constraint in mammalian genomic sequence." Genome research 15.7 (2005): 901-913.

[56] K. S. Pollard,M. J. Hubisz, K. R. Rosenbloom, and A. Siepel, ‘‘Detection of nonneutral substitution rates on mammalian phylogenies,’’ Genome Res., vol. 20, no. 1, (2010): pp. 110–121.

[57] W. James Kent et al., ‘‘The human genome browser at UCSC,’’ Genome Res., vol. 12, no. 6, (2002): pp. 996–1006.

[58] M. Kircher et al., ‘‘A general framework for estimating the relative pathogenicity of human genetic variants,’’ Nature Genetics, vol. 46, no. 3, (2014):pp. 310–315.

[59] A. Burga and B. Lehner, ‘‘Beyond genotype to phenotype: Why the phenotype of an individual cannot always be predicted from their genome sequence and the environment that they experience,’’ FEBS J., vol. 279, no. 20, (2012):pp. 3765–3775.

[60] J. R. Hart et al., ‘‘The butterfly effect in cancer: A single base mutation can remodel the cell,’’ Proc. Nat. Acad. Sci., vol. 112, no. 4, (2015): pp. 1131–1136.

[61] D. Devos and A. Valencia, ‘‘Practical limits of function prediction,’’ Proteins Struct. Funct. Genetics, vol. 41, no. 1, (2000):pp. 98–107.

[62] D. M. Fowler et al., ‘‘High-resolution mapping of protein sequence-function relationships,’’ Nature Methods, vol. 7, no. 9, (2010):pp. 741–746.

[63] A. Hannun et al., ‘‘Deep speech: Scaling up end-to-end speech recognition,’’ arXiv1412. 5567, 2014, pp. 1–12.

[64] R. Wu, S. Yan, Y. Shan, Q. Dang, and G. Sun, ‘‘Deep image: Scaling up image recognition,’’ (2015) arXiv1501.02876.

[65] J. Dean et al., ‘‘Large scale distributed deep networks,’’ in Proc. Adv. Neural Inf. Process. Syst., 2012, pp. 1223–1231.

[66] A. Agarwal, O. Chapelle, M. Dudı´k, and J. Langford, ‘‘A reliable effective terascale linear learning system,’’ J. Mach. Learn. Res., vol. 15, (2014):pp. 1111–1133.

[67] A. Coates, B. Huval, T. Wang, D. J. Wu, and A. Y. Ng, ‘‘Deep learning with COTS HPC systems,’’ in Proc. 30th Int. Conf. Mach. Learn., 2013, vol. 28, pp. 1337–1345.

[68] J. T. Dudley and A. J. Butte, ‘‘In silico research in the era of cloud computing,’’ Nature Biotechnol., vol. 28, no. 11, (2010):pp. 1181–1185.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[69] L. D. Stein, ‘‘The case for cloud computing in genome informatics,’’ Genome Biol., vol. 11, no. 5, (2010): p. 207.

[70] L. D. Stein, B. M. Knoppers, P. Campbell, G. Getz, and J. O. Korbel, ‘‘Data analysis: Create a cloud commons,’’ Nature, vol. 523, no. 7559, (2015):pp. 149–151.

[71] D. Sculley et al., ‘‘Machine learning: The high interest credit card of technical debt,’’ in Proc. SE4ML Softw. Eng. Mach. Learn., 2014, pp. 1–9.

[72] J. Shim et al., ‘‘Nanopore-based assay for detection of methylation in double-stranded DNA fragments,’’ ACS Nano, vol. 9, no. 1, (2015):pp. 290–300.

[73] H. Tilgner et al., ‘‘Accurate identification and analysis of human mRNA isoforms using deep long read sequencing,’’ G3 (Bethesda), vol. 3, no. 3, (2013):pp. 387–397.

[74] D. Sharon, H. Tilgner, F. Grubert, and M. Snyder, ‘‘A single-molecule long-read survey of the human transcriptome,’’ Nature Biotechnol., vol. 31, no. 11, (2013):pp. 1009–1014.

[75] B. Treutlein, O. Gokce, S. R. Quake, and T. C. Su¨dhof, ‘‘Cartography of neurexin alternative splicing mapped by single-molecule long-read mRNA sequencing,’’ Proc. Nat. Acad. Sci. USA, vol. 111, no. 13,(2014) pp. E1291–E1299.

[76] O. Stegle, S. A. Teichmann, and J. C. Marioni, ‘‘Computational and analytical challenges in single-cell transcriptomics,’’ Nature Rev. Genetics, vol. 16, no. 1, (2015):pp. 133–145.

[77] G. K. Marinov et al., ‘‘From single-cell to cell-pool transcriptomes: Stochasticity in gene expression and RNA splicing,’’ Genome Res., vol. 24, no. 3, (2014):pp. 496–510.

[78] S. S. Dey, L. Kester, B. Spanjaard, and A. Van, ‘‘integrated genome and transcriptome sequencing from the same cell,’’ Nature Biotechnol., vol. 33, no. 3, (2015):pp. 285–289.

[79] M. D. Ritchie, E. R. Holzinger, R. Li, S. A. Pendergrass, and D. Kim, ‘‘Methods of integrating data to uncover genotype-phenotype interactions,’’ Nature Rev. Genetics, vol. 16, no. 2, pp. 85–97, 2015.

[80] Liu, Hui, et al. "CADgene: a comprehensive database for coronary artery disease genes." Nucleic acids research 39.suppl 1 (2011): D991-D996.

[81] Guyon, Isabelle, and André Elisseeff. "An introduction to variable and feature selection." Journal of machine learning research 3.Mar (2003): 1157-1182.

[82] Budak, Hüseyin, and SemraErpolatTaşabat. "A Modified T-Score for Feature Selection." (2016): 845-852.

[83] Kummer, Olena, Jacques Savoy, and Rue Emile Argand. "Feature selection in sentiment analysis." (2012).

[84] Sahoo, PK-Riedel, and T. Mean Value Theorems. Functional Equations. World Scientific, 1998.

[85] http://www.sjsu.edu/faculty/gerstman/StatPrimer/t-table.pdf, 2017)

[86] Suykens, Johan AK, and Joos Vandewalle. "Least squares support vector machine classifiers." Neural processing letters 9.3 (1999): 293-300.

[87] Murphy, Kevin P. "Naive bayes classifiers." University of British Columbia (2006).

[88] An, Tae-Ki, and Moon-Hyun Kim. "A new diverse Adaboost classifier." Artificial Intelligence and Computational Intelligence (AICI), 2010 International Conference on. Vol. 1. IEEE, 2010.

[89] Kotsiantis, Sotiris B., I. Zaharakis, and P. Pintelas. "Supervised machine learning: A review of classification techniques." (2007): 3-24.

[90] Barrett, Tanya, et al. "NCBI GEO: archive for functional genomics data sets—update." Nucleic acids research 41.D1 (2013): D991-D995. (GEO: http://www.ncbi.nlm.nih.gov/geo/)

[91] Powers, David Martin. "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation." (2011).