Journal of
Computer Science & Systems Biology

Research Article Open Access

Novel Incremental Ranking Framework for Biomedical Data Analytics and Dimensionality Reduction: Big Data Challenges and Opportunities

Emad Elsebakhi[1]*, Ognian Asparouhov[2] and Rashid Al-Ali[1]

[1]*Biomedical Informatics Research Division, Sidra Medical and Research Canter, Doha 26999, Qatar*
[2]*ME Dai LexisNexis, Elsevier, 4901 Vineland Road, Orlando, Florida 32811, United States of America*

Abstract

Currently, due to the availability of massive biomedical data on each individual, both healthcare and life Science is becoming data-driven. The input-attributes are structured/un-structured data with many challenges, including sparse-binary attributes with imbalanced outcomes, non-unique distributed structure and high- dimensional data, which hamper efforts to make a clinical decision in clinical practice. In recent decades, considerable effort has been made toward overcoming most of these challenges, but still there is an essential need for significant improvements in this field, especially after integrating both omics and phenotype data for future personalized medicine. These challenges motivate us to use the state-of-the-art of big data analytics and large-scale machine learning frameworks to confront most of the challenges and provide proper clinical solutions to assess physicians in clinical practice at the bedside and subsequently provide high quality care while reducing its cost.

This research proposes a new recursive screening incremental ranking machine learning paradigm to empower the desired classifiers, especially for imbalanced training data, to create suitable data-driven clusters without prior information and later reduce the dimensionality of large biomedical data sets. The new framework combines many binary-attributes based on two criteria: (i) the minimum power value for each combination and (ii) the classification power of such a combination. Next, these sets of combined attributes are investigated by physicians to select the proper set of rules that make clinical sense and subsequently to use the result to empower the desired healthcare event (binary or multinomial target) at the bedside. After empowering the target class categories, we select the k-significant risk drivers with a suitable volume of data and high correlation to the desire outcome, and next, we establish the proper segmentation using AND-OR associative relationships. Finally, we use the propensity score to handle the imbalanced data, and next, we build break-through machine learning/data mining predictive models based on functional networks' maximum-likelihood and Newton-Raphson iterative matrix computation mechanism to expedite the implementations within high performance computing platforms, such as scalable MapReduce HDFS, Spark MLlib, and Google Sibyl. Comparative studies with both simulated and real-life biomedical databases are carried out for identifying specific biomedical and healthcare outcomes, such as asthma, breast cancer, gene mutations selection and genomic association studies for specific complex diseases. Results have shown that the proposed incremental learning scheme empower the new classifier with reliable and stable performance. The new classifier outperforms the current existing predictive models in both high quality outcome and less expensive in execution time, especially, with imbalanced and sparse with high-dimensional big biomedical data. We recommend future work to be conducted using real-life integrated clinic-genomic big data with genome-wide association studies for future personalized medicine.

Keywords: Biomedical data; Big data; Clustering; Personalized medicine; Parallelized distributed high performance computing; Large-scale machine learning; Classification; Data-driven; Electronic medical record; Next generation sequencing; Combinations and clinical data

Introduction

Due to the advances in both digital electronic medical record (EMR) systems and next generation sequencing (NGS); most hospitals and biomedical research institutes have massive amounts of biomedical data (size of peta-bytes of data); which pose challenges to be stored; visualized; and analysed. Therefore, there is an essential need for a breakthrough computational intelligence engine to query, analyse, and handle this large amount of biomedical data, at least to (i) improve population health and physicians' reliability in their daily clinical practice and (ii) improve the research discoveries within translational biomedical research and precision medicine. The main focus of this research is to develop or apply machine learning frameworks to handle the most common challenge problems in biomedicine and healthcare, especially after integrating both genotype and phenotype data. The advances in both big data analytics and large-scale machine learning paradigms is the key to success in designing future biomedical computational tools that are required to meet the challenges in healthcare [1,2].

Machine learning and knowledge discovery methods fall into four main categories: (i) supervised learning (forecasting and classification) (ii) unsupervised learning (clustering) (iii) semi-supervised learning and (iv) association rules with inductions. The specific challenge that we address in our research is the development of suitable decision support strategies based on novel innovative machine learning algorithms for the design and interpretation of novel predictive models for specific healthcare events using the available large data sets of individual or population biomedical data to fulfil the essential needs for personalized medicine at the bedside.

***Corresponding author:** Emad Elsebakhi, Biomedical Informatics Research Division, Sidra Medical and Research Canter, Doha 26999, Qatar, Tel: +00974-6698-3779; E-mail: eelsebakhy@sidra.org

Received May 12, 2015; **Accepted** June 10, 2015; **Published** June 12, 2015

Citation: Elsebakhi E, Asparouhov O, Al-Ali R (2015) Novel Incremental Ranking Framework for Biomedical Data Analytics and Dimensionality Reduction: Big Data Challenges and Opportunities. J Comput Sci Syst Biol 8: 203-214. doi:10.4172/jcsb.1000190

Recently, the advances in biomedical technology (EMR with NGS) have greatly increased the amount of available information that is often relevant to clinical decision support and cohorts of experts. These advances have the enormous potential of creating innovative data mining and machine learning algorithms toward improving diagnostic or prognostic accuracy as well as therapy selection capabilities. However; the available big data biomedical information comes with a larger risk of data overload and suboptimal utilization of the information in clinical practice, especially in precision medicine with biomarkers and drug discoveries and in predicting an individual's appropriate responses to a specific drug.

The motivation behind our research is to develop the capability of harnessing big data advanced analytics to manage the correct patients at the correct time with the correct interventions and with high quality treatments while reducing the costs. There are many real-life examples of the use of predictive modelling in both clinical practice and biomedical informatics which include (i) minimizing the mortality and re-admission rates of specific patients and providing real-time monitoring to provide assessments to physicians to save lives within the ICU, especially those preterm children and patients who are admitted with chronic diseases and (ii) determining the Single Nucleotide Polymorphisms (SNPs) loci in a genome that contains single-base variations across a population, which is essential in identifying diagnoses and predicting many complex diseases and common traits.

Related Work and Literature Review

It is known that clustering is the main pillar of machine learning in addressing un-supervised learning and similarities in multi-dimensional data. Although there are many statistical and computer science techniques for handling such issues, they are still used in some applications, but most of them have drawbacks, specifically, decision trees, association rules, apriori optimization, self-organized map, and fuzzy clustering methods. The scope of this research is to propose new methodology to overcome most of these drawbacks, especially within imbalanced and sparse large biomedical data sets. Next, we conduct comparative studies and summarize existing methods with their limitations in a literature review section, and with implementations, provide results interpretations and future outlooks.

With the current challenges in biomedical data, the computations in machine learning algorithms must be expedited. This step can be performed using the key of successful new parallel and distributed HPC computing using multiple processors (multi-cores) within Sidra Biomedical Informatics Service Hub HPC distributed systems, based on either batch or online sequential modeling; to perform enormous computations through a GPFS/LSF platform or MapReduce and Spark scalable performance model. It has been proven that both MapReduce and Spark platforms have many implemented algorithms such as, genetic algorithms and particle swarm optimization [3-5]. We can use both MapReduce and Spark with GPFS/LSF platforms to implement the new framework and the best classifier computation bottleneck for estimating the parameters of the model; which are complicated matrix computations with matrixes from very large training data sets. Therefore; we introduce a reliable solution for matrix multiplications, transposition, and inverses through two separate MapReduce functions.

Assume that we are given biomedical data D={(x_i, y_i); i=1, . . ., n} where y_i ε {(0,1,..., c-1) for p binary predictors belongs to {0;1}; which are sparse variables with some limitations to connect with the desired categorical target Y and $c \geq 2$. The focus is to develop a novel data mining algorithm to empower the desired group categorical target and overcome the difficulties in the building of significant relations/partitions to enhance future predictions in both healthcare and biomedical industry classification problems. Therefore, select the set of variables that is appropriate for participating in empowering the target class categories (for example, we select k input variables, (X_1; X_2;...; X_k); and then, we determine the appropriate associated interactions that empower the target class categories based on the most appropriate strategy of selecting this k-significant list of risk drivers with a high volume of data and strong relationship to the desired biomedical healthcare events.

The benefit of the desired recursive screening clustering data mining algorithm is to select the best combination of input features within the biomedical data and compute the classification power of a specific combination, and then, to create intermediate risk drivers within their own specific cluster using the proper combinations based on the AND; OR associative relationships. The main focus is to overcome the most common challenge problems in addressing big biomedical data, namely, sparse characteristics in a high-dimensional domain and addressing the curse of dimensionality. Therefore, we must find the appropriate significant interactions (combinations) of the given data from the available sparse biomedical variables that will be utilized to empower the outcome class categories. The last part of the problem is to design/develop a new data mining classifier based on constrained functional networks with a least squares and conjugate gradient optimization criterion. This new classifier considers both domain experts and data-driven knowledge discovery using a minimum description length to overcome both local optima and complexity due to the high dimensionality and sparse input feature space.

Challenges and Motivations

In working with biomedical data, we face many challenges, namely, sparse and high dimensionality with its corresponding collinearity and reduction problems. Therefore, the attribute contributions in identifying a specific healthcare event will be meaningless. In addition, due to the sparse within the input-space features; the predictive modelling computations will face challenging problems in determining the importance and meaning of the high volume of data within a specific cluster during the visualization and predictions. Furthermore, the expansion of the high-dimensional input biomedical feature space will lead to an ill-condition feature-space matrix, which leads to over fitting problems.

To assess and overcome some of the above-referenced challenges, we are proposing a new recursive screening clustering incremental ranking machine learning algorithm to handle biomedical data's common challenges. This new machine learning framework uses the priori structure information algorithm and association rule strategies to improve the classification or regression performance and then performs an assessment to identify the important and significant features (input attributes). The advantage of our new mathematical algorithms framework will expedite the implementation processes of biomedical data analytics and the predictions of both continuous and categorical healthcare outcomes, then it provides high quality treatment while reducing the cost.

To achieve our research, it is essential to review the state-of-the-art of the popular machine learning techniques that can work with sparse biomedical data. Based on the findings, we specify the most common drawbacks and limitations of these techniques with multi-dimensional sparse overlapping big biomedical data. We recommend future work to be conducted using simulated and real-life (integrated clinical and omics) data to check the performance and comparative studies with statistical techniques, decision trees, association rules, fuzzy clustering means, k-nearest neighbour, SOM and apriori algorithms.

Novel Recursive Clustering Algorithm

The desire training algorithm of the recursive screening clustering technique is to select the proper combination of predictors (input variables) from given biomedical data with multidimensional sparse binary attributes. This set of rules will be revised by the physicians and then used to enhance the classification power of the desired binary target. The entire implementation to obtain a specific combination is briefly summarized in algorithm 1:

Algorithm 1: New incremental recursive clustering

- **Step 1:** *Suppose that we have a binary classification problem with G_1 – (Group$_1$/Class$_1$) with n_1 observations and G_2 – (Group$_2$/ Class$_2$) with n_2 observations; where (n) or $N=n_1+n_2$. The goal is to create p binary predictors: x_1, x_2,… xi,… xp; (it is known that the binary/sparse biomedical data has some limitations in connecting with the desired target and that it is difficult to build significant relations to predict future healthcare outcomes). The goal is to find important interactions (partitions or clustering of data) between some of the sparse variables that will empower the desired target class/group.*
- **Step 2:** *Select the set of variables that is appropriate for participating (for example; we select k variables: x_1, x_2,… xi,… xk); and then select the best interactions (appropriate clusters) out of it. This step can be achieved by building an appropriate strategy for selecting this k-significant list of risk drivers with the appropriate high volume of data and a strong relationship with the available target (both continuous and multi-category values).*
- **Step 3**: The core idea of the recursive clustering mechanism is to select many parameters:
- *MinIntCnt=minimum number of variables to be included in each interaction (obviously 2<=MinIntCnt<=k);*
- *MaxIntCnt=maximum number of variables to be included in each interaction (obviously MinIntCnt<=MaxIntCnt<=k);*
- *Nmin=minimum number of observations for which a combination has the value 1. The traditional value of Nmin is 5 or 10 or 20);*
- *PowerMin=minimum power value for the combination (if Power ≥ PowerMin – save this combination); where Power ≥ 0.5 (as seen below);*
- *for count=MinIntCnt:MaxIntCnt;*

  select a new combination; (COMB(i)) with the following characteristics; for example:

(i) COMB(i)=1 if: x2=1; x3=1; x4=1; x6=1 (all combinations contains ONLY binary variables with value=1).

COMB(s)=AND(x2=1; x3=1; x4=1; x6=1);

Therefore; assuming that Ns is the total number of observations; where COMB(i)=1 (N_{s1} belongs to G_1 and N_{s2} belongs to g_2); where $Ns=N_{s1}+N_{s2}$. If Ns<Nmin; then skip this combination (exclude it from the calculations using genetic algorithms or simulated annealing to handle the sparse and multidimensional input-space and the curse of dimensionality, then reduce the number of combinations during the processing) – otherwise; the procedure is very time consuming.

(ii) Calculate the classification power of this combination (there are different formulas for its definition), for example;

(iii) Compute the following:

Power(COMB(i))=max (Ns1/Ns;Ns2/Ns);

Obviously, Power (any combination) ≥0.5; if Power≥ PowerMin; then save COMB(i).

(iv)Save the set of attributes and rules: The binary predictors that are included in this combination (all of them have the value 1); which is the value of Power (COMB (i)); will be saved.

- *Next; the physicians or case managers will investigate this set of rules and check their proper adequacy from a medical point of view. Based on both domain expert and data-driven conclusions; we keep only the rules that fall into common choices.*

(v)Dominating Group/Class will have the following values:

- *1 if Ns1>Ns2; (if G1 is the target class) and*
- *0 if Ns1 ≤ Ns2; (if G2 is the target class).*
- ***END.***

Initializations and critical cases

Here, we present the basic steps within the novel desired training algorithm using many unconditional and conditional loops to reduce the rules; for example; the probability of each group is important; due to the following details:

- We may have different minimum power values (precision) for each class, specifically for imbalanced data (*very rare*), (*G=1 in 98%; and G=0 in 2%*) of the observations. Therefore, we use two different *PowerMin values*; such as *PowerMin(1)=99.5 and PowerMin(0)=25%* (we are interested in such a combination only because it is a very rare class) or
- We may check the threshold in Step 2 and then select the *Target Class (in this case G=0)* and give the *PowerMin* value only for the *Target Class.*
- The two above steps must be performed after computing the classifier power using the following formula:

Power (COMB (i))=Ns_TargetClass/Ns.

Based on this novel recursive clustering paradigm we will be able to handle a sparse and multidimensional input space and the curse of dimensionality. In addition, save these binary predictors within each combination. Next, the physicians or case managers will investigate this set of rules and check their adequacy from a medical point of view.

According to both the domain expert and data-driven conclusions, we keep only the rules that fall into common choices. During the implementations, we used the most popular data analytics and modeling quality measures and requirements to discover the risk drivers and to be certain of both the efficiency and reliability of the developed new framework using both simulated and real-life practical issues in both biomedicine and healthcare domains, in other words;

If the target is continuous values:

- For the given set of attributes within a set of data; D; we compute the following quality measures:

- Compute the most common statistical summaries (volume of positive values; average; median; and standard deviation);
- Do the transformation (1/attribute; (attribute)^2; sqrt(attribute); log(1/(1+attribute));
- Compute the correlation between the target and actual attribute and the transformations in step 2;
- Apply flag criteria to select the higher correlation within step 3;

- Sort the data based on both the volume of positive values and the values within the flag criterion column;
- Select the attributes that have the higher correlations.

If the target is in the multinomial/binary category:

- For the given set of attributes within a set of data; we compute the following quality measures:
    - Compute the common statistical summaries (volume of positive values; ratio in each group);
    - Compute the F-value within each category;
    - Do the transformation (1/attribute; (attribute)^2; sqrt(attribute); log(1/(1+attribute));
    - Compute the correlation between the target with the actual attribute and each transformation within step 2;
    - Apply flag criteria to select the higher correlation within step 3;
- Sort the data based on the volume of positive values and the values within the flag criterion column;
- Select the attributes that have the higher correlations.

Clustering and associations/matching: Similarities with handling the biomedical sparse data and dimensionality reduction.

Advantage of the new clustering framework

The advantages of the new clustering technique can be summarized as follows:

- The proposed novel recursive clustering paradigm can handle a sparse and multidimensional input space and the curse of dimensionality then, reduce the number of combinations during the development of the data mining predictive modeling classifier;
- The binary predictors within each combination are investigated by the physicians or case managers and are checked for adequacy from a medical point of view. Therefore; based on both the domain expert and data-driven conclusions;
- We keep only the ones that fall into common choices;
- The new paradigm can be deployed within parallelized distributed HPC systems to handle big biomedical data and fulfill the need of personalized medicine;
- The novel data mining recursive clustering technique is simple and has O(n) computations with no prior information/probability similar to the ones with the current techniques;
- In our study; we use a new data mining classifier that is based on constrained functional networks with a least squares and conjugate gradient optimization criterion. This new classifier considers both the domain expert and data-driven knowledge discovery using a minimum description length to overcome both local optima and complexity due to high dimensionality and a sparse input feature space;
- With the new framework; there is no need to worry about the number of clusters; k; and no sensitivity to outliers that leads to skewed means or to initial conditions that produce different results of the clustering;
- The new paradigm is unlike a k-means cluster technique; but we have no non-linear ratio-scale. Therefore; the new paradigm has ***O(n);*** whereas; the k-means cluster computation has a complexity of ***O(K*n*No. Iterations);***
- The new algorithm uses both associations that are based on volume of each class category and its quotation power classification support; unlike the Apriori algorithm *O (2*No. of items)*. In addition; the Apriori algorithm is a time-consuming algorithm; especially when k is large;
- In the new recursive clustering algorithm; there is no need to specify the distance functions among the clusters, unlike Fuzzy C-Means and k-means clustering with the Mahalanobis distance with a mean and covariance matrix, which are very sensitive to outliers and centers;
- The new algorithm is optimal due to the use of a genetic algorithm and minimum description length, which leads to fast local optima, unlike the ID3, which is not optimal, due to its use of expected entropy reduction;
- The new algorithm does not suffer from any problems when we build the rules, unlike the decision trees, which suffer from the problem of errors propagating throughout the tree, which is a very serious problem as the number of classes' increases.
- Therefore, we have developed the appropriate significant interactions (combinations) of a given data set from the available sparse biomedical variables that will be utilized to empower the outcome class categories. The next sections will include the entire process and implementations of the proposed novel recursive and incremental algorithm to empower functional networks classifier using numerous of clinical and genomics data and simulation studies within high performance computing platform: scalable MapReduce HDFS, Spark MLlib, and Google Sibyl [6-8].

Novel Machine Learning Classifier

Background

In the past few years; functional network models have become popular frameworks for the predictive modeling of different real-life applications, such as medicine and business. The obtained results have proven that functional networks can be considered to be a remarkable data mining knowledge discovery paradigm for predicting both continuous and categorical outcomes [9-11]. However, this new intelligence system framework has not been utilized in the biomedicine and healthcare industries, especially for big data. The motivation behind this research is to propose machine learning based on functional networks and use the strength of the MapReduce and Spark distributed platforms to expedite the computations of functional networks that are based on maximum-likelihood estimation (FunNets- MLE) classifiers and then design a suitable decision and overcome the common challenges within big biomedical data and fulfill the requirements of personalized medicine.

In FunNets-MLE, the MLE is a good concave optimal solution that is certainly convergent but is very sensitive to noisy data, additionally, it takes an enormous number of iterative computations to approximate. The Newton-Raphson method is a fast iterative process for approximation, especially within a distributed file systems (DSF) platform; such as Hadoop or Spark, to calculate matrix transposes; computations; and inverses; especially with sparse and multi-dimensional data. In this research; we are utilizing the minimum description length criterion that take care of both outliers/missing values and handle collinearity among attributes; then the RSVD methodology can work fast with no obstacles. In addition; we use different criteria for imputation to deal with missing data, depending on the percentage of missing values within each attributes; for instance; use mean/median of the rest of non-missing values within each attribute and assign a high value when missing; then this will prevent a record with a missing value to be integrated in the neighbourhood or significant of attributes.

Iterative Newton-Raphson functional networks classifier

The core of Newton's technique is to guess the initial root; **X***i*.; to estimate the new value of the root; $\mathrm{X}_{i.}+1\ \mathrm{F}(\mathrm{X}_{i.-1},\Theta_k)$; in other words;

$$\mathrm{X}_{i.}+1=\mathrm{X}_{i.}-\mathrm{F}(\mathrm{X}_{i.},\Theta_k)\left(\frac{\partial\mathrm{F}(\mathrm{X}_{i.},\Theta_k)}{\partial\Theta_k}\right)^{-1} \tag{1}$$

Where $|\in_s|=|(\mathrm{X}_{i.}+1)^{-1}(\mathrm{X}_{i.+1}-\mathrm{X}_{i.})|\times100$ is an error, say, 10^{-4}. If $|\in_s|>10^{-4}$ then continue; otherwise, stop.

Given the data D={Y, **X**}, $\mathrm{X}\in R^P, Y\subseteq R$ and $\mathbf{X}_{i.}=(x_{i1},....,x_{ip})$ Consider the models; $\pi_{ik}=P(y_i=k\,|\,\mathrm{X}_{i.})=f(x_{i1},....,x_{ip})+\varepsilon_i$ for k=1, ... , c; and i=1, . . . , n. The goal is to find π_{ik}; which is the probability that observation i falls in class A_k. Assume that $\pi_{ik}=p(\tau_{ik})$ where p (•) is a probability function, which must satisfy the probability conditions; and τ_{ik} is the desired model:

$$\tau_{ik}=g_k(x_{i.},\Theta_k)+\varepsilon_{ik} \tag{2}$$

Here, $\tau_{ik}=\log_e\left[n_k\left(n-\sum_{k=1}^{c-1}n_k\right)^{-1}\right]$ and $\pi_{ik}=e^{g_k(x_i,\Theta_k)}\left(1+\sum_{k=1}^{c-1}e^{g_k(x_i,\Theta_k)}\right)^{-1}$

then; the goal is to estimate Θ_k. The function p(•) can be sigmoidal; logit; probit and Cumulative Distribution Function (CDF). During the implementation process in our research; the function that we are interested in applying Newton-Raphson to, for the purpose of expediting the computations, is the first derivative of the Log-Likelihood function;

$$l(\boldsymbol{\theta}_{jk})=\sum_{i=1}^{n}\left\{\sum_{k=0}^{n}\left[y_{ik}\log_e(\pi_{ik})\right]\right\} \tag{3}$$

where y_{ik} is the coded matrix; then; the desired Newton-Raphson's is written as in [7].

$$\Theta_{k,new}=\Theta_{k,old}\left(\frac{\partial^2 l(x_{i.},\Theta_k)}{\partial\Theta_k\partial\Theta_k^T}\right)^{-1}\left(\frac{\partial l(x_{i.},\Theta_k)}{\partial\Theta_k}\right). \tag{4}$$

Therefore; by choosing $\psi_{11kl}(x_{i1})=\left\{1,x_{i1},...x_{i1}^m\right\}$ with a constant term for $g_{11k}(x_{i1})$ and the family $\psi_{rjkh}(x_{ij})=\left\{x_{ij},...x_{ij}^m\right\}$ without the constant term for the other $g_{rjk}(x_{ij})$ we will obtain the following form:

$$\Theta_{k,new}=\Theta_{k,old}-\left[\left(\mathbf{X}^T\mathrm{M}\,\mathbf{X}\right)^{-1}\right]\left[\mathbf{X}^T(Y-\pi_k)\right] \tag{5}$$

Where,

$$\mathrm{M}(i,i)=\pi_k\left(\mathbf{X}_{i.},\Theta_{k,old}\right)\prod_{k=1}^{c-1}\pi_k\left(\mathbf{X}_{i.},\Theta_{k,old}\right);$$

for j, l=0,1,. . . ., $(m+1)^P$ and k, k_1=0, 1, c-1. We note that if the process is initialized at zero and the MLE is known to have concave shape properties, then convergence is guaranteed. Therefore, the implementation processes turn out to be matrix multiplications and inverses:

$$\left[\left(\mathbf{X}^T\mathrm{M}\,\mathbf{X}\right)^{-1}\right]\text{and}\left[\mathbf{X}^T(Y-\pi_k)\right].$$

To avoid the enormous computational time, we use one of the available parallel distributed HPC platforms; namely; the Hadoop MapReduce framework, or Spark, or GPFS/LSF systems. Therefore; we will easily perform the mapping and reducing computations and specify the key value outputs/inputs of each of the Map and Reduce functions; as accomplished in [3]. The parallel-FunNets-MLE architecture can be drawn easily; which expresses both the first derivatives of the log-Maximum likelihood function and its corresponding *Jacobian* or the *Hessian matrix*; $\left(\frac{\partial^2 l(x_{i.},\Theta_k)}{\partial\Theta_k\partial\Theta_k^T}\right)^{-1}$. The processes and training algorithm of FunNets-MLE is based on Newton-Raphson's technique according to equations 1 through 5.

Randomized large-scale singular value decomposition

To minimize the computational time of our algorithm; we utilize the scalable randomized large-scale singular value decomposition (RLSVD) within the available parallel distributed HPC platforms: 16 and 32 cores with the Hadoop MapReduce framework and Spark or GPFS/LSF systems. This novel randomized algorithm is a scalable singular value decomposition for optimizing the performance of the desire machine learning algorithm (Functional networks with Maximum-Likelihood estimations based on Newton-Raphson iterative approximations) [12]. The core idea is to quickly calculate the inverse of a matrix ***A*** or determine a rank-k approximation for an m-by-n matrix; ***A***; for large m and n (10^6 or more). Therefore; the goal is to find a rank of (k<<n); where $A=U\sum V^T$ Σ is a k-by-k non-zero-covariance matrix with $\sigma_1\geq\sigma_2\geq...\geq\sigma_k\geq0$ while U and V are orthonormal singular matrices of A.

The advantage of the RSVD method is in presenting a fast and easy way to solve problems using Eigen-value decomposition and finding spanning columns or rows of a given matrix and then minimizing the computational cost. The Matlab script for RSVD can be written as follows:

```
function [U; S; V]=rsvd(A; k)

[m; n]=size(A); nz=nnz(A);

P=randn(n; k + 5);

Y=full( A * P );

[Q; R]=qr(Y; 0);

B=full( Q' * A );
```

[Uhat; S; V]=svd(B; ‘econ’);

U=Q * Uhat;

U=U(:;1:k); S=S(1:k;1:k); V=V(:;1:k);

This approach offers excellent methodology for computing matrixes within a multi-processor and big biomedical data and then rapidly reducing the time of the computations; with information being gathered at an even faster pace.

U=randn(50*10000; 1000); V=randn(50*100000; 1000);

A=U*V’; [v; i; j]=find(A);[~;ids]=sort(v); ids=[ids(1:length(v)/500); ids(end-length(v)/500:end)]; A=sparse(v(ids); i(ids); j(ids));

for k=2:1000; tic; [U0; S0; V0]=svds(A; k); toc; end for k=2:1000; tic; [Ur; Sr; Vr=rsvd(A; k); toc; end

This type of methodology can be tested by simulating random data with large-scale dimensionalities; say 5×10^5 by 5×10^6; and for 500 attributes and having a non-zero covariance matrix; then; we have an almost 1000-times reduction in the computational time by using RSVD versus the common original SVDs in Matlab. The new RSVD elapsed computational time is 1.09 seconds; while the regular existing SVDs method in Matlab has elapsed computational time of 998.9 seconds. The greatest use for such a methodology in the future is; for example; biomedical informatics of next generation sequencing and whole-genome and DNA sequencing or calculating the associations and similarities within genome-wide association studies and with respect to common or rare diseases. Furthermore; this approach can be applied to determine the Principal Component Analysis using an empirical covariance matrix from some collection of statistical data and then to compute the singular value decomposition of the given matrix to find the directions of maximal variance.

Implementation and Discussion

The goal is to deploy the developed machine learning methodology based on functional networks maximum likelihood classifier with the Newton-Raphson iterative algorithm for many simulated and real-life biomedical and healthcare industry data. The results with the desired statistical quality measures for both regression and support vector machines are summarized in the tables and figures.

Data acquisition

Congestive heart failure (CHF) data: This clinical data consists of 4,310 observations and 199 input variables. The target of these data involves both groups: Control individuals with 3,288 (76.1%) patients and case individuals with 1,031 (23.9%) patients. The data were combined together to form classification biomedical data of 4,319 individuals with a precision of 23.9%; which is a small percentage (imbalanced data) and represents a challenge for any machine learning or data mining classifier. The entire de-identified data and its ownership was supported by MEDai Inc; Elsevier Company, Orlando FL, USA.

Implementations and discussion

All of the implementations were accomplished using Mahout in the Hadoop DFS MapReduce platform and MLlib Spark system. The new script of the novel recursive clustering framework was implemented in R and Matlab using 32- and 16-core HPC distributed systems. The implementations were performed based on the data-in-hand of both real-life congestive heart failure (CHF) data, very imbalanced breast cancer data; genomic wide association study data and gene mutations to identify children severe asthma exacerbation; and simulated studies.

Congestive heart failure data: The utilization of the new recursive screening clustering mechanism on this clinical data has been achieved; and at the end; we list the significant risk drivers (the high-risk variables) and all of the processes in the tables and graphs. The entire set of processes and implementations are as follows:

Step 1: Split for (a) Babies+pregnant women; and (b) all others (Everything below is for all others).

Step 2: Run the proper statistical inference pre-processing calculations as explained above for classification problems with binary categorical outcomes. The desired target (dependent variable) is 0/1; where 1 is the target class; (for example; Readmission; Mortality; ICU admission; Sepsis). Therefore; we consider the treatment cost and length of stay (LOS) to be continuous outcomes (regression problem with the other script code: LOS ≥ 10 and LOS ≥ 25. Day 1 model truncation: LOS=max 25; the goal is ≥ 10 (red); and therefore; the PPV is the N1 criterion; not R^2.

Step 3: Either for continuous or categorical outcomes; select the top 10 measures with the max ratio (in general F-ratio ≥ 4 is a good value for a high risk) and (volume of clustering data: *N>20 or 10*); if $N=G_0+G_1$ and G_1 (*target class observations*) ≥ *5.*

- *Select these 10 splits or drivers; paste them and run (R-Script for either Classification or Regression)_again; select the first one and paste its results in the final list: High-Risk variables (X’s=Readmission/Mortality; ICU; or Sepsis);*
- *Split according to the first clustering; and run (R-Script for either Classification or Regression) the script for the other nine splits or risk drivers;*
- *Select the first one (max ratio) if the ratio is still good and is not correlated with the first selected; paste it in High- Risk (X’s); and so on;*
- *Suppose that after the fourth selected variable (out of 10); the other six’s ratio decreases significantly.*
- *Let us say that the first four selected measures are X1; X2; X3; and X4. Therefore; go to all of the other nodes and split them as follows: OR(X1; X2; X3; X4).*
- *Go to the second node (where X1=X2=X3=X4=0 or none of them is positive);*
- *Go to step 2.*

END

After repeating this procedure several times, we will come up with a list of High-Risk (*X’s*) variables. Suppose that all of the measures (input variables) that left have a ratio of <4.

- Now; select these with a ratio of >3 or >2 and a high volume (N=several thousands). Assume that Y is such a measure. Therefore; we split Y and go to Step 1, in this way, we search for interactions: (*Y>0)(Z>0*) that have a ratio of ≥ 4.
- The list of high-risk measures will contain thirty to seventy variables and interactions; along with their statistics.

We note that any continuous variable can be transformed into several binary variables by dividing it into intervals or following some binary code structures. For example; the variable age can be written as follows:

- Age (0-5) years: (binary); it is equal to 1 if Age belongs to the interval (0,5); and 0 otherwise;
- Age (5-10) years: (binary); it is equal to 1 if Age belongs to the interval (5,10); and 0 otherwise;
- Age (10-25) years: (binary); it is equal to 1 if Age belongs to the interval (10,25); and 0 otherwise;
- Age (25-45) years: (binary); it is equal to 1 if Age belongs to the interval (25,45); and 0 otherwise;
- Age (45-65) years: (binary); it is equal to 1 if Age belongs to the interval (45,65); and 0 otherwise;
- Age (65-100) years: (binary); it is equal to 1 if Age belongs to the interval (65,100); and 0 otherwise.

Next, the physicians or case managers revised the obtained list of significant variables; and then; the results produced the desired High-Risk Cluster. The same procedures are performed for the first two clusters: (i) Babies; and (ii) Pregnant women. An example is the high risk variable for sepsis (High-Risk-Sepsis).

To determine the scoring performance of the new clustering paradigm and build a suitable risk score; we calculate all of the combinations of size's two; three; four; and five attributes out of the available 199 input attributes; which sums up to $\sum_{u=2}^{5}\binom{199}{u}$ combinations for which each comprise a set of attributes. In reality; this number is enormous and would involve expensive computations. The new recursive incremental algorithm reduces the time of the computations of these combinations to only 0.06% and produces a reasonable number for the minimum number of combinations (29+68+112+74); which is 283 sets of combinations instead of the above enormous number of combinations; comprising (630+7140+58,905+376,992)=443,667 sets of rules.

For simplicity; we propose only 20 rules out of the obtained list of rules in Table 1, using different combinations of at least five predictors to empower the classifier; which considers the physician's point of view with the data-driven measures as well. These sets of rules were built using the new recursive screening clustering approach. In addition; there were other measures that were included in this list, but they were correlated with measures that were already selected with higher ratio values; for example, two, three, and four predictors within different combinations of the risk factor out of the 198 input attributes in high-dimensional clinical data of congestive heart failure databases; as is shown in Table 1. Moreover, the area under the curve (ROC) for the average performance of the functional network predictive models that are based on the fast Newton-Raphson computational techniques with over 1000 runs is presented in the curve in Figure 1; which represents the specificity versus sensitivity for the congestive heart failure disease.

Breast cancer surveillance consortium (BSCS) imbalanced data: To test the performance of the new predictive modeling framework on an imbalanced data with less than 10% (positive class or case-data has only 10% or less of the whole given observations). This can be achieved for "Breast Cancer Surveillance Consortium (BCSC) data: http://breastscreening.cancer.gov/rfdataset/. The data contains approximately; 2.4 million screenings mammograms and associated self-administered questionnaires. The primary goal of these studies was not readability; but rather highest risk detection performances and impact levels of each risk factor: our goal is to provide a risk level without making the decision (breast cancer or not) in place of the physician. The data originally contained 2,392,998 records of index screening mammograms from women included in the Breast Cancer Surveillance Consortium (BSCS) [13-14]. Among the 2,392,998 records of the database; 9314 cases of invasive breast cancer were diagnosed in the first year of follow up. Here; we are facing highly imbalanced data with a positive class accounting for only 0.39% of all records. We observe that; the BCSC database contains most of the known breast cancer personal factors. It is the largest database publicly available that includes breast density information.

As pointed out by [14-15] "information and dialog with more patient involvement in the decision-making process" are key words in

Different combinations of five attributes	P0=G0/n0	P1=G1/n1	P0/P1	P1/P0	Q	Y
(HighRiskRenal; ChRenF_All; AcuRenF_All; Nephritis; MI_All)	0.00090	0.00970	0.09	10.63	10.63	1
(HighRiskRenal; ChRenF_All; AcuRenF_All; Nephritis; OldMIOnly)	0.00150	0.01450	0.1	9.57	9.57	1
(HighRiskRenal; ChRenF_All; AcuRenF_All; Nephritis; IschHrtDis)	0.00150	0.01070	0.14	7.02	7.02	1
(HighRiskRenal; ChRenF_All; AcuRenF_All; Nephritis; PulmHrtDis)	0.00150	0.00870	0.17	5.74	5.74	1
(HighRiskRenal; ChRenF_All; AcuRenF_All; Nephritis; Cardmyopty)	0.00360	0.02040	0.18	5.58	5.58	1
(HighRiskRenal; ChRenF_All; AcuRenF_All; Nephritis; CABGprev)	0.00180	0.01450	0.13	7.97	7.97	1
(HighRiskRenal; ChRenF_All; AcuRenF_All; Nephritis; Angina)	0.00300	0.01550	0.2	5.1	5.1	1
(HighRiskRenal; ChRenF_All; AcuRenF_All; Nephritis; AtrialFib)	0.00360	0.01840	0.2	5.05	5.05	1
(HighRiskRenal; ChRenF_All; AcuRenF_All; Nephritis; Arrhythmia)	0.00520	0.03100	0.17	6	6	1
(HighRiskRenal; ChRenF_All; AcuRenF_All; Nephritis; ValveDis)	0.00360	0.02130	0.17	5.85	5.85	1
(HighRiskRenal; ChRenF_All; AcuRenF_All; Nephritis; HighRiskResp)	0.00580	0.03880	0.15	6.71	6.71	1
(HighRiskRenal; ChRenF_All; AcuRenF_All; Nephritis; HighRiskMSkel)	0.00150	0.01070	0.14	7.02	7.02	1
(HighRiskRenal; ChRenF_All; AcuRenF_All; Nephritis; Diabetes_All)	0.00730	0.04070	0.18	5.58	5.58	1
(HighRiskRenal; ChRenF_All; AcuRenF_All; Nephritis; HighRiskDiab)	0.00610	0.03490	0.17	5.74	5.74	1
(HighRiskRenal; ChRenF_All; AcuRenF_All; Nephritis; HighRiskNeuro)	0.00490	0.02720	0.18	5.58	5.58	1
(HighRiskRenal; ChRenF_All; AcuRenF_All; Nephritis; HighRiskGILiv)	0.00240	0.01940	0.13	7.97	7.97	1
(HighRiskRenal; ChRenF_All; AcuRenF_All; Nephritis; HighRiskInfect)	0.00180	0.01260	0.14	6.91	6.91	1
(HighRiskRenal; ChRenF_All; AcuRenF_All; Nephritis; HighRiskBlood)	0.00490	0.03200	0.15	6.58	6.58	1
(ChRenF_All; AcuRenF_All; Nephritis; RenInsuff; MI_All)	0.00090	0.00780	0.12	8.5	8.5	1
(ChRenF_All; AcuRenF_All; Nephritis; RenInsuff; OldMIOnly)	0.00120	0.01070	0.11	8.77	8.77	1

Table 1: Congestive heart failure (CHF) data: The 20 rules of combinations of five predictors out of 75 obtained rules.

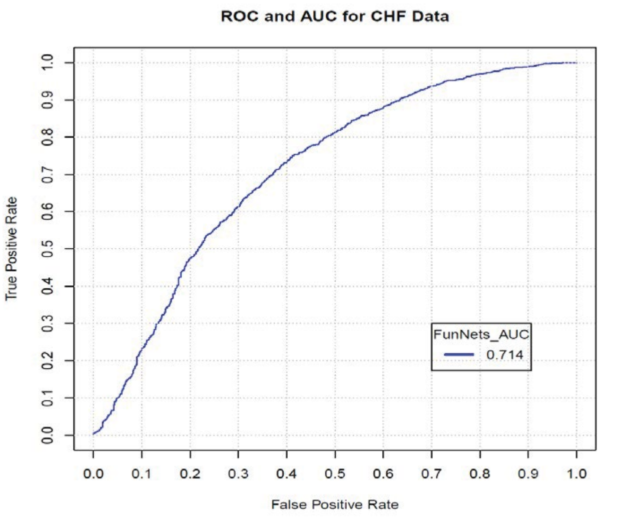


Figure 1: CHF Data: ROC Curve for functional networks classifier over 1000 runs.

dealing with cancer. Therefore a major challenge in the field of medical counselling is to assess clinicians and radiologists with adequate tools to help them to assess their patients' breast cancer risk and to show easily how risk factors impact global risk. Yet, the risk scores that were built upon statistical models were not adopted at bedside regardless their accuracy. This is due to the end-users of these tools are neither oncologists nor clinicians and underlying too difficult to use during a medical consultation. Hence; it is essential to develop a new risk score reliable technique to be used within the current medical decision process.

To build a risk score that helps to detect highest risk profiles among general population; the mining algorithms has to provide a risk value without labelling a woman profile [14] and therein references. Dealing with general population means we are facing highly imbalanced data with a breast cancer incidence rate lower than 1000 new cases for 100,000 women. Dealing with such imbalanced data can be done at both algorithmic and data levels. At data level by choosing a different cost or rebalancing positives or negatives examples. At algorithmic level; it is possible to make a K-nearest-neighbour algorithm more sensitive to the minority class by modifying the neighbourhood boundaries or by using a class confidence weight to handle imbalanced data during the labelling step. The authors in [14] implemented logistic regression; neural networks; k-nearest neighbour with different values of k; and decision trees risk models were built with 4 to 10 risk factors depending on the menopausal status. Comparative studies were carried and they reported that the obtained area under ROC curve results were in between 0.631and 0.642 for premenopausal/postmenopausal women.

In our study; we have done our best to handle both missing values and outliers using different criteria and then overcome the ill-conditions and collinearity within the attributes in feature-space. In addition; we follow different imputation criterion to deal with missing data; depending on the percentage of missing values within each attributes; for instance; use mean/median of the rest of non-missing values within each attribute and we follow the same strategies that were shown in [14] by assign a high value when missing; then this will prevent a record with a missing value to be integrated in the neighborhood or significant of attributes. Therefore, we have cleaning data with 181,903 observations and 13 sparse attributes (describing various pathological and mammography characteristics of the women); while the target (indicating diagnosis of breast cancer within one year of the screening mammogram). The summary statistics of the entire features of the BCSC data are shown in Table 2; where G0+=Summation of non-zero values supported "control group"; G0%=G0+/actual total number of control group; G1+=Summation of non-zero values supported "case group"; G1%=G1+/actual total number of Case group. These input variables have been determined to influence a woman's risk for developing breast cancer and will henceforth be referred as risk factors. We observe that the given data is a very imbalanced data; where it has 175,629 individuals without breast cancer diseases (control data); which is 96.55% of the available complete observations; while it has only 6,274 individuals with breast cancer diseases (case data); which is representing only 3.45% (a very small percentage compare to the control subset) of the available complete observations. As it is shown in [14]; the goal is to use this breast cancer data (for two cohort groups: premenopausal and postmenopausal women) and then develop novel predictive models to (i) present the dependencies between breast cancer risk factors and then (ii) how to have accurate breast cancer risk score.

The authors in [14] created their own scoring Performances as an experiment set that was designed to test how the k-nearest-neighbour algorithm perform on the BCSC data (choosing k=5). As one of their constraints is to build a readable risk score; they select all combinations with a size s of 1 to 6 attributes among n=12 available attributes; meaning they have in total 2,509 combinations to test. A first way of assessing results of these combinations is to look at the best combinations by size. These results are obtained using an Euclidian space using a 2-norm Euclidian distance as they are not significantly better; when improved; using another p-norm measures. In Gauthier et al.; the authors find that the first list of all possible combinations (from 1 to 6 attributes) [14] performs better than the two specialized regression models obtained on pre- and postmenopausal women by Barlow WE [13] with an AUC of 0.642; based on *agegrp; density; brstproc*; *lastmamm* combination. In addition; they use the domain knowledge expert and they recorded the top 15 performance results before and after expert advice.

As before; we run the new algorithm to build a suitable risk score to empower the classifier for better prediction; by calculating all of the combinations of size: two to 9 attributes out of the available 12 input attributes, which sums up to $\sum_{u=2}^{9}\binom{12}{u}$ combinations for which each comprise a set of attributes. In reality; this number is enormous

Index	Variable name	G0+	G1+	G0%	G1%	Ratio
1	menopaus	147125	5236	83.77	83.46	1
2	agegrp	175629	6274	100	100	1
3	density	175629	6274	100	100	1
4	race	175629	6274	100	100	1
5	Hispanic	77106	1727	43.9	27.53	0.63
6	bmi	175629	6274	100	100	1
7	nrelbc	66595	2073	37.92	33.04	0.87
8	brstproc	68697	2220	39.11	35.38	0.9
9	lastmamm	76854	2564	43.76	40.87	0.93
10	surgmeno	119062	4061	67.79	64.73	0.95
11	hrt	112720	4147	64.18	66.1	1.03
12	invasive	0	4912	0	78.29	25

Table 2: Diagnosis of invasive carcinoma in situ breast cancer within one year of the index screening mammogram

and would involve expensive computations. The new recursive incremental algorithm reduces the time of the computations of these combinations and produces a reasonable number for the minimum number of combinations (3+8+11+14+9+12+10+5); which is 72 sets of combinations that can be easily interpreted by physicians and use it to empower the desire classifier; then build a proper clinical decision system. These set of rules are very easy to interpret and smaller than the actual rules and it can be instead of (66+220+495+792+924+495+220) =4,004 comprising sets of rules. For the sake of space and simplicity; we are shown we are shown in Table 3; the top 5 rules of the combinations of 7 and 9 attributes with its corresponding AUC values.

We observe that the new framework for sure will outperform logistic regression; neural networks; and decision trees according to the published results within both [13-14]. In addition; the ROC/AUC show improvement around 24%; which means that, we may save lives of 1,506 patients from those 6,274 patients. Therefore; if each one of these individuals is spending $50,000 a year; then; we reduce the total cost of more than $75 million a year. In addition; minimize their future complications; minimize their LOS; ICU; and minimize both lengths of stay and/or emergency room visits with Readmission rates. In addition; the area under the curve (ROC)/AUC is 0.795 for the average performance of the novel method over 1000 runs is presented in the curve in Figure 2, which represents the specificity versus sensitivity for the congestive heart failure disease. Furthermore, we can develop novel large-scale predictive models to present the dependencies between breast cancer risk factors and have accurate breast cancer risk score that is easy to be used by physicians or case managers.

SNP selection and Genomic Wide Association Study Data in Identifying Children Asthma Exacerbation: By utilizing the real-life genomic wide association study (GWAS) data from [16] to forecast asthma exacerbations in children using random forests classifiers. According to the National Institute of Allergy and Infectious Disease report in (2015): https://www.aafa.org/display.cfm?sub=42&id=8; it is known that asthma is a complex disease known to be influenced by genetic; clinic; and environmental factors: There are 26.7 million individuals or about 9.7% of the U.S. populations have had asthma during their lifetime. In the year 2000, asthma exacerbations resulted in 1,499 deaths; 1.1 million hospital days, and $2.9 billion in direct expenditures in the U.S. There are more than 3,600 deaths due to asthma each year, many of which are avoidable with proper treatment and care. It is the leading chronic disease among children and it is the third-ranking cause of hospitalization in children. Approximately 25.9 million Americans suffer from asthma (8% of adults; 10% of children); women account for almost 65% of asthma deaths; and asthma affects over 230 million people worldwide. The prevalence of asthma has been increasing since the early 1980s across all age; sex and racial groups. Each year; asthma accounts for more than 14 million doctor visits; almost 2 million emergency room visits; 439,000 hospitalizations (average length of stay (LOS) for asthma is 3.6 days); and more than 3,600 deaths. The annual cost of asthma is estimated to be $56 billion. The direct costs accounted for nearly $50.1 billion (hospitalizations the single largest portion of direct cost) and indirect costs of $5.9 billion (lost earnings due to illness or death).

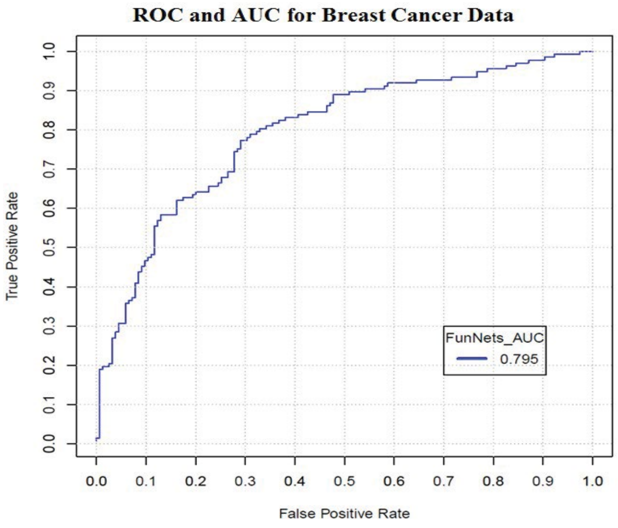


Figure 2: CBSC Data: ROC Curve for increment recursive functional networks classifier over 1000 runs.

Recently; the authors in [16] predicted the children severe asthma exacerbations using ensemble learning based on “Random Forests” (RF), Bagging and Boosting algorithm using several clinical factors including the forced expiratory volume in one second as a percent of predicted (FEV 1%); oral corticosteroid usage; age; and gender (sex). However; these factors by themselves are limited in their ability to successfully predict severe asthma exacerbations. Therefore; they explore the potential power of a multi-single nucleotide polymorphism (SNP) model and GWAS data as incorporated into RF together with clinical relevant risk factors to effectively predict such complex diseases; this algorithm is applied to the prediction of exacerbations in a population of childhood asthmatics participating in the Childhood Asthma Management Program (CAMP): Stage 1 (population of 127 exacerbation cases and 290 non-exacerbation controls) and Stage 2 (population of 50 exacerbation cases and 114 non-exacerbation controls). However; the accuracy limitations; sparse attributes; and the selection of the significant SNP mutations and find both similarity and associations with complex diseases are still challenges. Therefore, to meet these challenges and be able to provide proper personalized medicine for every individual, it is essential to use the state-of-the-art-of large-scale machine learning predictive modeling with big data analytics to predict the children severe asthma exacerbations using integrated genomic-clinic and environmental big biomedical factors. This kind of prediction would therefore have direct prognostic significance and might form the basis for the development of novel therapeutic interventions.

According to the SNP data from genome-wide scans that are available through the National Institutes of Health (NIH) database of Genotypes and Phenotypes (dbGaP) and CAMP parents participations as it is shown in [16] the data acquisition and the primary outcomes of the problem-in-hand can be summarized as follows:

- **Clinical covariates:** Age; gender; pre-bronchodilator FEV 1%; and treatment group (clinical traits) are known to be associated with asthma exacerbations: Age and pre-bronchodilator FEV 1%; are coded as numeric variables; gender/sex is coded as 1 for male; 2 for female; treatment group is coded as 1; 2; 3 for three different treatments.

- **Primary outcome:** The occurrence of either an emergency room visit or a hospitalization for asthma symptoms at any time during the clinical trial period was used to define a severe asthma exacerbation.

GWAS data: Of the CAMP participants; 417 Caucasian parent-trios-children were genotyped using the Infinium II HumanHap550v3

Genotyping BeadChip, and 164 Caucasian non-trio cohort children were subsequently genotyped using the Human660W-Quad BeadChip. Over 500,000 SNPs were successfully genotyped in the CAMP trios; with a reproducibility of >99.99%. Reproducibility is based on 4 samples that were each genotyped 15 times in the experiment. According to the explanations within [16] the Genotype quality is validated using the Mendel option of PLINK v0.99r http://pngu.mgh.harvard.edu/purcell/plink; [17] verifying allele calls against Ref-Seq to ensure correct orientation; and testing for extreme departures from Hardy Weinburg equilbrium in the parents.

Selection of SNPs: Focusing on the trio probands as their initial test population; they used RF importance scores to rank and select SNPs in two steps. At each step; they used SNPs as predictors to predict asthma exacerbations with RF; and obtained the RF importance score of each of the SNPs. At the first step; the computed RF importance scores for all SNPs genome-wide; 4,000 at a time; in chromosomal order. At the second step; they ranked all SNPs based on their RF importance scores; selected the top 4,000 SNPs; and re-run RF with these selected SNPs to re-rank them.

Prediction model building with RF: The 417 Caucasian trio samples (Stage 1 samples) were genotyped before the 164 cohort samples (Stage 2 samples), and were used to build and train the RF models to predict asthma exacerbations. The R package random forest version 4.5-25 was used to build RF models in this study: http://cran.r- project.org/web/packages/randomForest/index.html. The RF predicted score is the percentage of trees voting for "yes". During this step and the steps described in "selection of SNPs" above; RF parameter "ntree" (number of trees to grow) were set to be 1,500 - a relatively large number to ensure stable prediction results, and all other parameters, including mtry, were set to use the default values.

The authors in [16] used emergency room visits or hospitalizations as the definition of a severe asthma exacerbation; they first identified a list of top GWAS-SNPs ranked by Random Forests importance score for the CAMP of 127 exacerbation cases and 290 non-exacerbation controls. They predict severe asthma exacerbations using the top 10 to 320 SNPs together with age; gender; pre-bronchodilator FEV 1 percentage predicted; and treatment group. The authors in [16] found that the Area Under the Curve (AUC) score of 0.54 using the clinical traits alone; suggesting the phenotype is affected by genetic as well as environmental factors. On the other hand; they test/validate the predictions in an independent set of the CAMP population shows that severe asthma exacerbations can be predicted with an AUC=0.66 with 160-320 SNPs in comparison to an AUC score of 0.57 with 10 SNPs. Their study shows that a random forests algorithm can effectively extract and use the information contained in a small number of samples. In addition; random forests; and other machine learning tools, can be used with GWAS studies to integrate large numbers of predictors simultaneously.

We follow the same procedures as of [16] in order to assess the performance of the new classifier based on the integrated phenotypic and omics data with the environmental factors (417 CAMP asthmatic family-trio children genotyped and 164 CAMP asthmatic family-trio children genotyped) using stratified sampling methodology of random partition 70% (training) and 30% (testing) with the repetition of computations for 1000 times and one of the common benchmark stopping criterion. In addition; we use the selected clinical traits and SNPs as predictors with the two types of controls (i) permutation control; (ii) other random SNP control; and (iii) the black demarcation separates the top 4,000 SNPs (important and significant towards the target) from the rest (not significant towards the target). The entire population stages were utilized for both training and validation of the new parallelized functional networks classifier to identify the children severe asthma exacerbations and compute the common statistical quality measures; namely; sensitivity and specificity; with the Receiver Operating Characteristic curve (ROC curve) and the Area under the ROC Curve (AUC).

To gain the usefulness of random-forest ensemble learning with the new classifier; we note that the importance of random forest score measures the relative contribution of a predictor to the desire target; which is similar to our set of combinations of rules and apply the propensity score of each of the SNPs and plot it in chromosomal order. By drawing the demarcation separates to identify the top 4000 SNPs with the highest significant score to the target.

To compare the performance of the new classifier with the ensemble learning random forest classifier; we implement both incremental clustering-set of combination rules and parallelized functional networks classifier using the following schemas: (i) The clinical data (age, gender, pre-bronchodilator FEV 1%, and treatment group) alone; and then (ii) similar to [16] we use the integrated repository data (age; gender; pre-bronchodilator FEV 1%, and treatment group; different numbers of SNPs selected based on the propensity score based on the significant SNPs predictors; then at the end we are able to empower the parallelized functional networks classifier to identify the children severe asthma exacerbations with varying degrees of success. We carried over during the implementation processes; the minimum description length criterion that take care of both outliers/missing values and handle collinearity among attributes, then the RSVD methodology can work fast with no obstacles.

We concluded that the new classifier with the support of the set of combined rules add trust and reliable to the classifier and increase its AUC values based on the nonlinear relationship with the available phenotypic attributes with demographic predictors is 0.62, which is better than the corresponding values in the existing RF values. On the other hand, by adding of the 10 significant SNPs as it was done in the case of RF, we found that the exacerbations increased the AUC to 0.65, which is better than 0.57. Furthermore; by adding more important SNPs columns to increase the predictability of asthma exacerbations; we got the independently replicated AUC values of of 0.67, 0.71, and 0.71 for 40, 160 and 320 SNPs, which is 7.5% improvement, respectively compared to the RF values within [16]. We conclude that the new parallelized functional networks classifier empowered through the novel criterion of set of combination rules according to propensity score for the important SNPs and we are able to identify the children asthma exacerbations through the use of hundreds of significant SNPs in a novel large-scale machine learning predictive modeling based on functional networks and minimum description length. The comparative studies versus the achieved ensemble learning and random forests model by [16] were reasonable and improve the classification accuracy, which has impact on each individuals treatment cost. In addition, it will improve the healthcare outcomes, for instance, minimize the length of stay, minimize re-admission and emergency room visits; and can be used for disease intervention profiling. Furthermore, as it was expressed in [16], "the new classification model" can increase our understanding of the biologic mechanisms behind why only certain individuals with asthma are at risk for exacerbations; as well as the basis for the epistatic (gene-gene) interactions underlying asthma severity; providing insight into novel preventative and therapeutic strategies".

Simulated mixture data: Classification and time of the

computations: Investigating the strengths and capabilities of the new technique requires both real-life data and simulation studies using a 32- core machine. The simulated data are from mixture bivariate predictors and the binomial response of two groups; where each group consists of a mixture of two subclasses (groups); C$_1$ and C$_2$, as follows:

$$C_1 = \{x : x \in N(\mu_{11}, \Sigma_{11}) \cup N(\mu_{12}, \Sigma_{12})\};$$

$$C_2 = \{x : x \in N(\mu_{21}, \Sigma_{21}) \cup N(\mu_{22}, \Sigma_{22})\};$$

Where,

$$\begin{matrix} \mu_{11} = [3,3]; \mu_{12} = [5,5]; \\ \mu_{21} = [5,3]; \mu_{22} = [3,5]; \end{matrix} \text{ and } \Sigma_{ij} = \begin{pmatrix} 1 & \pm 0.5 \\ \pm 0.5 & 1 \end{pmatrix}; \textit{for } i, j = 1, 2$$

Figure 3 shows the scatter plot; and there is overlap between the two classes; C$_1$ and C$_2$. A mixture of the two normal distributions for each class is supposed to make the decision boundary more complex.

We consider different training sets and different testing sets according to the two-tenth-fold stratified criterion to measure the performance of many classifiers; including functional networks (FunNets), with a maximum likelihood (MLE) based on the iterative Newton-Raphson's method. We summarized the results in the tables and graphs in Figure 4 and Table 3.

Based on the obtained results; we conclude that the comparative studies among these classifiers over the 1000 runs can assess in identifying the desire target in a proper stable decision. We find that the new functional networks classifier with the empowering combination sets of rules is having the lowest time of execution with the highest average correct classification rate, and the lowest misclassification error. On the other hand; the support vector machine is the second highest in both correct classification rate classifier and misclassification error; but it has execution time more than logistic regression, this is because kernel function and high-dimensional input-space. Moreover, the neural networks have the highest execution time; due to its trial and errors and randomly chosen initial weights and its architecture is very complex; but it has reasonable performance with the third place in accuracy. The logistic regression in this simulated study has the worst performance with reasonable execution time because its fast linear approximations that leads to less reliable performance. Therefore, overall the new functional network outperforms the most common techniques, namely; logistic regression, neural networks, and support vector machines; it is faster than all classifiers and has reliable and efficient performance (Table 4).

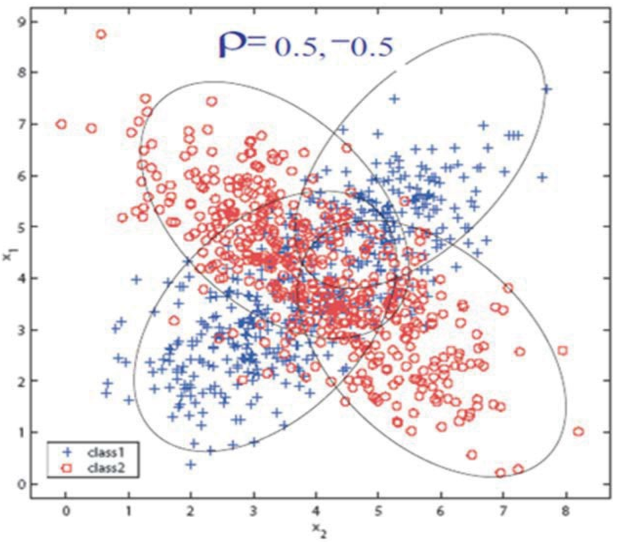
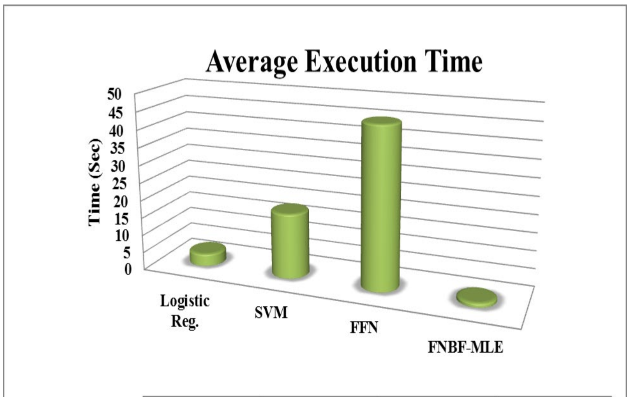


Figure 3: Simulated binary data from the mixed binomial categories in 2D with ρ ± 0.5.



	Logistic Reg.	SVM	FFN	FNBF-MLE
🟩 Avg. Exec. Time	3.519	18.392	45.151	1.204

Figure 4: Time of Computations within 1000 runs.

Different combinations of five attributes	AUC
agegrp; density; bmi; race; nrelbc; lastmamm; hispanic; brstproc; surgmeno; hrt	0.795
agegrp; density; bmi; nrelbc; lastmamm; hispanic; brstproc; surgmeno; hrt	0.77
agegrp; density; race; bmi; nrelbc; lastmamm; brstproc; surgmeno; hrt	0.74
agegrp; density; bmi; nrelbc; lastmamm; brstproc; surgmeno; hrt	0.705
agegrp; surgmeno; lastmamm; density; brstproc; hrt; bmi	0.695

Table 3: BCSC data: The Top 5 rules of combinations of five predictors out of 72 obtained rules.

Conclusions and Future Outlook

The new algorithm is optimal due to the use of the genetic algorithm and minimum description length, which leads to fast local optima, unlike the ID3 method, which is not optimal, due to its use of expected entropy reduction. In addition, the new algorithm does not suffer from any problems when we build rules; unlike the decision trees, which suffer from the problem of errors propagating throughout the tree, which is a very serious problem as the number of classes' increases. Moreover, the propensity score handled the imbalance, which empowers the functional network classifiers.

As is shown from the empirical simulated data; we observed that the randomized singular value decomposition out-performs the regular singular value decomposition for computing matrixes within multi-processors and big biomedical data and rapidly reduces the time of the computations with an almost 1000-times reduction; and the information is gathered at an even faster pace.

- We observe that the benefit of having the Newton-Raphson iterative matrix computation was in expediting the quality and performance of the recursive and modeling processes using the HPC and scalable HDFS MapReduce and Spark MLlib. The implementation processes are initialized at zero with a starting approximation and then use the concave shape properties of

Classifier	Avg. Exec. Time	Avg. CCR	Avg. ASCE
LogReg.	3.519	0.502	23.591
SVM	18.392	0.719	8.821
FFN	45.151	0.684	13.019
FunNets	**1.204**	**0.815**	**3.281**

Table 4: Average over 1000 runs of four classifiers.

MLE; the convergence is guaranteed, which determines the computations for only the matrix multiplications, transpose, and inverse. Furthermore, the new recursive incremental algorithm reduced the times of the computations of these combinations to only 0.06% and produces a reasonable number for the minimum number of combinations, which are 283 sets of combinations instead of 443,667 sets of rules. These small numbers for the number of sets of rules will be easy to retrieve or investigate by either physicians or case managers for specific patients and fast to use in clinical practice at the bedside.

- For the imbalanced breast cancer data, we observe that the ROC/AUC show improvement around 24%, which means that; we may save lives of 1,506 patients from those 6,274 patients. Therefore, if each one of these individuals is spending $50,000 a year, then; we reduce the total cost of more than $75 million a year. In addition, minimize their future complications, minimize their LOS, ICU, and minimize both lengths of stay and/or emergency room visits with read mission rates. Furthermore; we can develop novel large-scale predictive models to present the dependencies between breast cancer risk factors and have accurate breast cancer risk score that is easy to be used by physicians or case managers.
- Based on the two utilized biomedical applications that are used for the cases of breast cancer and congestive heart failure, the results show that the new frameworks have a reliable performance with a high impact on empowering the classifier and building a minimum number of rules that is easy to be revised by physicians and fast to use in clinical practice. We recommend future work to be conducted using simulated and real-life data to compare the performance of the new paradigm with the existing techniques.
- We investigate the strength and capabilities of the new classifier on imbalanced biomedical big SNP selection and genomic wide association study data, we conclude that the new classifier with the support of the set of combined rules add trust and reliable to the classifier and increase its AUC values based on the nonlinear relationship with the available phenotypic attributes with demographic predictors. In addition, we find that the exacerbations increased the AUC to 0.65. Furthermore; by adding more important SNPs columns to increase the predictability of asthma exacerbations; we got the independently replicated AUC values of 0.67, 0.71 and 0.71 for 40, 160 and 320 SNPs; which is 7.5% improvement better than the existing quality measures in literature. Therefore; we recommend the new classification model for more complex biomedical data to increase our understanding of the biologic mechanisms behind why only certain individuals have specific complex disease, as well as the basis for the epistatic (gene-gene) interactions underlying asthma severity; providing insight into novel preventative and therapeutic strategies".

Acknowledgment

The authors acknowledge the support of Sidra Medical and Research Canter; Doha; Qatar for the facilities utilized to perform the present work and for their support. They are also grateful to MEDai Inc. (now LexisNexis); Orlando; Florida; USA for their support and implementations.

References

1. Molla M, Waddell M, Page D, Shavlik J (2004) Using machine learning to design and interpret gene- expression microarrays. AI Magazine 25: 23-44.
2. Cristianini N, Hahn M (2006) Introduction to Computational Genomics. Cambridge, UK: Cambridge University Press.
3. Zhen Liu, Meng Liu (2011) Logistic regression parameter estimation based on Parallel matrix computation. Communications in Computer and Information Science 164: 268-275.
4. Singh S, Kubica J, Larsen S, Sorokina D (2009) Parallel Large Scale Feature Selection for Logistic Regression. In: 9th SIAM International Conference on Data Mining, pp. 1165-1176.
5. McNabb AW, Monson CK, Seppi KD (2007) Parallel PSO using MapReduce. In: IEEE Congress on Evolutionary Computation, pp. 7-14.
6. Zaharia M, Chowdhury M, Das T, Dave A, Ma J. McCauley, et al. (2012) Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing in Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation April 25-27, 2012.
7. Zhuang Y, Chin WC, Juan YC, Lin CJ (2014) Distributed Newton method for regularized logistic regression. Lecture Notes in Computer Science 9078: 690-703.
8. Chu CT, Kim SK, Lin Y, Yu YY, Bradski G (2006) Map-Reduce for machine learning on multicore. In Proceed of Advances in Neural Information Processing Systems, pp. 281-288.
9. Enrique CF, Cobo A, Gómez-Nesterkin R, Ali HS (2000) A general framework for functional networks. Networks 35: 70-82.
10. El-Sebakhy EA, Hadi AS, Faisal KA (2007) Iterative least squares functional networks classifier. IEEE Trans Neural Netw 18: 844-850.
11. El-Sebakhy E, Asparouhov O, Abdulraheem A, Al-Majed A, Wu D, et al. (2012) Functional networks as a new data mining predictive paradigm to predict permeability in a carbonate reservoir. Expert Syst Appl 39: 10359- 10375.
12. Halko N, Martinsson PG, Tropp J (2009) Finding structure with randomness: Stochastic algorithms for constructing approximate matrix decompositions. Survey and Review section 53: 217-288.
13. Barlow WE, White E, Ballard-Barbash R, Vacek PM, Titus-Ernstoff L, et al. (2006) Prospective breast cancer risk prediction model for women undergoing screening mammography. J Natl Cancer Inst 98: 1204-1214.
14. Émilien G, Laurent B, Philippe L, Stéphane R (2015) A Nearest Neighbor Approach to Build a Readable Risk Score for Breast Cancer. Real World Data Mining Applications, 17: 249-269.
15. Testard-Vaillant P (2010) The war on cancer. CNRS Int Mag 17: 18-21.
16. Xu M, Tantisira KG, Wu A, Litonjua AA, Chu JH, et al. (2011) Genome Wide Association Study to predict severe asthma exacerbations in children using random forests classifiers. BMC Med Genet 12: 90.
17. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81: 559-575.