Hindawi
Complexity
Volume 2017, Article ID 7120691, 11 pages
https://doi.org/10.1155/2017/7120691

WILEY | Hindawi

Research Article

Sparse Learning of the Disease Severity Score for High-Dimensional Data

Ivan Stojkovic[1,2] and Zoran Obradovic[2]

[1]*Signals and Systems Department, School of Electrical Engineering, University of Belgrade, Bulevar Kralja Aleksandra 73, 11120 Belgrade, Serbia*
[2]*Center for Data Analytics and Biomedical Informatics, College of Science and Technology, Temple University, 1925 North 12th Street, Philadelphia, PA 19122, USA*

Correspondence should be addressed to Zoran Obradovic; zoran.obradovic@temple.edu

Received 11 May 2017; Revised 6 November 2017; Accepted 27 November 2017; Published 18 December 2017

Academic Editor: Sergio Gómez

Learning disease severity scores automatically from collected measurements may aid in the quality of both healthcare and scientific understanding. Some steps in that direction have been taken and machine learning algorithms for extracting scoring functions from data have been proposed. Given the rapid increase in both quantity and diversity of data measured and stored, the large amount of information is becoming one of the challenges for learning algorithms. In this work, we investigated the direction of the problem where the dimensionality of measured variables is large. Learning the severity score in such cases brings the issue of which of measured features are relevant. We have proposed a novel approach by combining desirable properties of existing formulations, which compares favorably to alternatives in accuracy and especially in the robustness of the learned scoring function. The proposed formulation has a nonsmooth penalty that induces sparsity. This problem is solved by addressing a dual formulation which is smooth and allows an efficient optimization. The proposed approach might be used as an effective and reliable tool for both scoring function learning and biomarker discovery, as demonstrated by identifying a stable set of genes related to influenza symptoms' severity, which are enriched in immune-related processes.

1. Introduction

Diseases and other health conditions require continuous monitoring and assessment of the subject's state. The severity of the condition needs to be quantified, such that it can be used to guide medical decisions and allow appropriate and timely interventions. Disease severity scoring functions are typically used to quantify a patient's condition. However, disease severity and health are often difficult to quantify. That is because they are essentially latent concepts and are not directly accessible or observable. In an absence of a direct measurement of health, the severity of a condition is estimated based on values of some surrogate variables that are observable and hopefully informative about the condition of interest. In clinical practice, commonly tracked variables include temperature, heart rate, blood pressure, and responsiveness, to name a few out of a myriad of possible other variables. A severity score is subsequently calculated from such observable quantities using some heuristic rules. Prominent examples of such rules are SOFA [1] score for sepsis, or more general ICU scoring systems like APACHE II [2]. Both relevant variables and associated heuristic rules are established in a consensus of expert bodies and relevant institutions based on experience and current understanding of a condition. That process is long and tedious and often results in extensively coarse scoring rules and a nonoptimal set of relevant observable variables.

Although utilizing data was always part of this process, recently it was acknowledged that it might be improved/complemented by using machine learning methods that can automatically extract both rules and relevant variables directly from the data. There are already a number of approaches for automatic learning of severity scores/rules from data. One way is to use discrete class labels

for building classifiers and subsequently use the probability of a sample belonging to a certain class as a quantification measure of severity [3]. Another supervised approach is to learn the severity score function in a regression manner [4] from some surrogate of severity highly associated with undesired outcomes down the stream. A downside is that it already requires a good candidate for scoring function. An additional issue is that it might be sensitive to censoring due to treatment, where the severity of state is not acknowledged because treatment prevented undesired outcomes from happening. Some of the mentioned drawbacks are addressed in a more recent approach [5, 6] that is based on a clever observation that comparing two cases according to severity is easier to assess than to directly quantify the severity of a particular case. It was built upon the existing work on learning scoring functions for information retrieval tasks [7]. However, even this approach might be inappropriate in some cases, since it learns the severity score as a function of all measured variables, which will affect its performance when there are irrelevant features or when the number of features is much higher than the number of samples [8]. In essence, features unrelated to severity will be present even in small sets of measured variables, and, in high-throughput measurements like gene expression, this might be an even larger obstacle.

In this paper, we present an approach to the problem of learning disease severity scores in presence of irrelevant or high-dimensional measurements. We build on top of existing efforts by simultaneously performing feature selections that are most relevant for severity score learning. In particular, we are introducing the L_1 norm in the formulation of ranking SVM [7] along with the temporal smoothness constraint [6]. Attractive regularization properties of L_1 norm are already well acknowledged and exploited in a number of statistical learning methods since its introduction [9]. The proposed formulation of sparse severity score learning forces weights of (most of) the features to be exactly zero, therefore effectively performing feature selection by learning the sparse linear scoring function. This novel severity score objective function is convex and nonsmooth and it precludes the direct use of convenient optimization tools like gradient-based methods. Therefore, in this work, we are also providing the reformulation of the problem into its dual that is smooth and that allows efficient optimization. Other than learning the severity score from the data, which is an important instrument for assessing severity, the methodology may also be used to discover the most relevant variables/features for the disease severity phenotype. Such findings might be further used to suggest novel (testable) hypotheses about causal relations leading to disease manifestation and also to inspire novel therapeutic approaches.

The rest of the article is structured as follows: Methods begins with the introduction to related work and continues with new formulation and derivation of its solution. Results begins with evaluation on intuitive synthetic examples where the advantages of sparse severity score framework over the nonsparse one are apparent. Results continues with the assessment on a gene expression dataset of H3N2 viral infection responses. Efficacy and the robustness of the proposed method are compared favorably against multiple alternative methods. Results is concluded with gene ontology overrepresentation analysis of the discovered subset of genes most useful for the scoring function.

2. Methods

2.1. Previous and Related Work. As mentioned in Introduction, some of the first proposed severity score learning methods are supervised approaches that solve classification or regression tasks and whose solution provides a way to calculate a severity score.

For example, in [4] Alzheimer's Disease severity, as measured by cognitive scores, was modeled as (temporal) multitask regression using the fused sparse group lasso approach. The approach was more concerned with the progression of the disease, hence the multitask formulation. However, as we are mostly interested in severity score mapping from a single time-point set of measurements, here we are presenting its more influential ancestor, the LASSO model [9]:

$$\underset{w}{\operatorname{argmin}} \quad \mathrm{LASSO}(w) = \frac{1}{2} \|Y - Xw\|_2^2 + \lambda \|w\|_1 . \tag{1}$$

Here, Y is column vector of n given numeric scores, associated with d dimensional measurement matrix $X_{n\times d}$, while w denotes the solution in form of a d-dimensional column weight vector. We will use this model as one of the baselines for comparison as it is one of the main workhorses of biomarker selection [10] and even statistical learning in general.

Another approach used sparsity-inducing L_1 norm in combination with classical loss function for learning disease severity scoring function [3]. They proposed using L_1 regularized Logistic Regression model (among others), to model the severity scores for the abnormality of the skull in craniosynostosis cases:

$$\begin{aligned} \underset{w}{\operatorname{argmin}} \quad & L_1\mathrm{LogReg}(w) \\ & = \sum_{i=1}^{n} \log\left(1 + \exp\left(-Y_i\left(X_i w\right)\right)\right) + \lambda \|w\|_1 . \end{aligned} \tag{2}$$

This Sparse Logistic Regression formulation is another related model, as it also results in a sparse vector of feature weights w that essentially regress the decision boundary between the severity classes and might be used as a mapping function for severity scores. In (2), $Y_i \in \{-1, 1\}$ is a binary label for ith row of data matrix X.

As outlined previously, these forms of supervision where estimates of severity score functions (or severity classes) are needed might be hard to obtain in order to be utilized for training the severity score automatically. On the other hand, obtaining the pairs of comparisons is an easier task. Seminal work of learning the scoring functions from the comparison labels is proposed in [7]. In that work, the ranking SVM formulation (see (3)) is developed to learn better document retrieval from click-through data. This great insight came from noticing that the clicked links automatically have greater

ranks compared to the ones not clicked. And such kind of data is much more abundant than the user provided rankings.

$$\begin{aligned}\underset{w}{\operatorname{argmin}} \quad & \operatorname{rankingSVM}(w) \\ &= \frac{1}{2}\|w\|_2^2 \\ &\quad + c\sum_{\{p,q\}\in O}\max\left(0, 1-\left(X_p - X_q\right)w\right).\end{aligned} \tag{3}$$

Set O is composed of comparison of ordered pairs $\{p, q\}$, where p has a higher rank than q and which corresponds to rows of measurement matrix X_p and X_q, respectively. More recently the approach was adopted for learning the Sepsis Disease Severity Score [5]. In it (see (4)), the constraint that scoring function should gradually evolve over the time was introduced and hence a temporal smoothness term is added. In addition, nonsmooth Hinge loss ($\max(0, 1 - Xw)$) is replaced with its smooth approximation, Huber loss (L_h), to obtain the formulation of (linear) Disease Severity Score Learning (DSSL) framework:

$$\begin{aligned}\underset{w}{\operatorname{argmin}} \quad & \operatorname{DSSL}(w) \\ &= \frac{1}{2}\|w\|_2^2 + c\sum_{\{p,q\}\in O} L_h\left(1-\left(X_p - X_q\right)w\right) \\ &\quad + b\sum_{\{i,i+1\}_s\in S}\left(\frac{\left(X_{i+1}^s - X_i^s\right)w}{\left(t_{i+1}^s - t_i^s\right)}\right)^2.\end{aligned} \tag{4}$$

Temporal smoothness term in (4) penalizes high rates of change in severity in consecutive time steps t_i and t_{i+1} of a single subject s. Set of all consecutive pairs in all subjects is denoted by S and constants c and b are hyperparameters determining the cost of respective loss terms.

DSSL framework was adopted and extended in different ways. A multitask DSSL was proposed in [11], which utilizes matrix norm regularization to couple multiple distinct tasks. Nonlinear version of DSSL framework, as well as its solution in form of gradient boosted regression trees, was also proposed in [6]. Nevertheless, mentioned DSSL approaches are dense in a sense that they operate on all variables (in case of a linear version, all coefficients are typically nonzero). The approach in [11] is based on expensive proximal gradient optimization algorithm, which makes it unsuitable for high-dimensional problems. The utility of the approaches in [6] was presented on an application with a moderately small number of different pieces of clinical information, vitals, and laboratory analysis variables and it is not clear how the approach would perform in situations with high-dimensional data common in high-throughput techniques like genetic, genomic, epigenetic, proteomic, and so on.

Yet, high-throughput data is also a very rich source of useful biomarkers that could be used for diagnostic and prognostic purposes, as well as for obtaining insight into causal relations [12]. Therefore we are proposing an approach that is able to learn a (temporally smooth) scoring function from comparison data while simultaneously performing the selection of most relevant (important) variables.

2.2. Proposed Model Formulation. In our Sparse Learning of Disease Severity Score (SLDSS) formulation, we combine attractive properties (and terms) of previously mentioned approaches, ranking SVM (see (3)) [7], temporal smoothness constraint (see (4)) [6], and L_1 norm from sparse methods (see (1) and (2)) [3, 9]:

$$\begin{aligned}\min_{w} \quad & \operatorname{SLDSS}(w) \\ &= \frac{1}{2}\|w\|_2^2 + c\sum_{\{p,q\}\in O}\max\left(0, 1-\left(X_p - X_q\right)w\right) \\ &\quad + b\sum_{\{i,i+1\}_s\in S}\left(\frac{\left(X_{i+1}^s - X_i^s\right)w}{\left(t_{i+1}^s - t_i^s\right)}\right)^2 + \lambda\|w\|_1.\end{aligned} \tag{5}$$

In fact, since the model imposes both L_1 and L_2 norms on the feature vector w, it resembles the elastic net regularization [13], which has an advantage of achieving higher stability with respect to random sampling [14].

The solution w^* of the optimization objective defined in (5) serves as a sparse linear function $f(X) = Xw^*$ that may be applied on measurements from the new patient, to obtain a scalar value of severity that might be compared to previously assessed cases and inform further actions. The sparse vector w^* may also serve as an indicator of which features are the most influential for pairwise comparison. The formulation contains two nonsmooth terms, L_1 and Hinge loss, and therefore it is not directly solvable using off-the-shelf gradient methods. In DSSL formulation, the (nondifferentiable) Hinge loss is approximated with twice differentiable Huber loss, thus making the optimization criterion solvable using the second-order gradient methods (e.g., Newton and Quasi-Newton). In order to provide an efficient solution for the proposed nonsmooth objective, we will solve the smooth dual problem instead of relying on smooth approximation or nonsmooth optimization tools.

First we rewrite (5) into a more suitable form for which we will later provide the smooth dual problem. We aggregate the differences of measurements into single data matrix $D_{k\times d}$, where k is a number of pairs in the comparison set O. Similarly, we express measurement and temporal difference ratios as matrix $R_{l\times d}$, where rows are $R_i = (X_{i+1}^s - X_i^s)/(t_{i+1}^s - t_i^s)$ and l is a number of pairs in the consecutive measurements set S. We aggregate the L_2 norm and temporal smoothness terms (they are essentially weighting the square of optimization parameters) into a single weighted quadratic term $(1/2)w^TQw$, where $Q = I + 2bR^TR$, I being d-dimensional identity matrix. The first two terms, weighted quadratic norm and Hinge loss, resemble the well-known SVM criterion function that we will rewrite in its "soft" form with additional slack variables z_i and their associated constraints. Additional set of "dummy variables" y is introduced in L_1 term, with trivial constraints $w = y$. The equation of the rewritten SLDSS now reads

$$\begin{aligned}\min_{w} \quad & \operatorname{SLDSS}(w, z, y) = \frac{1}{2}w^TQw + c\sum_{i=1}^{k} z_i + \lambda\|y\|_1 \\ \text{s.t.} \quad & D_i w \geq 1 - z_i,\end{aligned}$$

$$z_i \geq 0,$$
$$\forall i \in \{1, \ldots, k\},$$
$$w = y. \tag{6}$$

Now we turn this constrained problem with inequalities and equalities into its Lagrangian dual. Constraints are moved to the criterion function as penal terms weighted by Lagrangian multipliers α, β, and γ. The equation of the SLDSS dual problem is

$$\begin{aligned} \min_{w,y,z\geq 0} \max_{\alpha\geq 0,\beta\geq 0} \quad & \text{Dual}(w, y, z, \alpha, \beta, \gamma) \\ & = \frac{1}{2} w^T Q w + c\mathbf{1}^T z + \alpha^T (\mathbf{1} - z - Dw) \\ & \quad - \beta^T z + \lambda \|y\|_1 + \gamma^T (w - y). \end{aligned} \tag{7}$$

Given that optimization criterion is convex and feasible (Slater's condition holds [15]), strong duality allows switching the order of maximization and minimization in (7), and minimization in primal variables can be safely performed first. Now we analyze the expression according to primal variables w, y, and z and find the minimizing conditions for each of them.

The dual formulation is the quadratic function of parameters w and we can find its optimal form as a function of new free parameters introduced in dual (by equating its gradient with zero):

$$\begin{aligned} \min_{w} \quad & \text{DUAL}(w) = \min_{w} \frac{1}{2} \left(w^T Q - \alpha^T D + \gamma^T \right) w \\ & \nabla_w \text{DUAL}(w) = w^T Q - \alpha^T D + \gamma^T = \mathbf{0} \\ & \Downarrow \\ & w^* = Q^{-1} \left(\alpha^T D - \gamma^T \right). \end{aligned} \tag{8}$$

Similarly, the expression for slack variables z is a linear combination of dual variables and it is minimal when the directional gradient is equated to zero vector, giving the optimality condition in a form of an equality constraint:

$$\begin{aligned} \min_{z} \quad & \text{DUAL}(z) = \min_{z} \frac{1}{2} \left(c\mathbf{1}^T - \alpha^T - \beta^T \right) z \\ & \nabla_z \text{DUAL}(z) = c\mathbf{1}^T - \alpha^T - \beta^T = \mathbf{0} \\ & \Downarrow \\ & \beta = c\mathbf{1} - \alpha. \end{aligned} \tag{9}$$

Resulting equality constraint $\beta = c\mathbf{1} - \alpha$ in combination with inequality $\beta \geq \mathbf{0}$ can be reduced to just one constraint $\alpha \leq c\mathbf{1}$, which removes β from further consideration.

For minimization over dummy variables y, we use the convex (Fenchel) conjugate function of the expression [15] and obtain optimality condition as inequality constraint over the infinity norm of the dual variable:

$$\begin{aligned} \min_{y} \quad & \text{DUAL}(y) = \min_{y} \lambda \|y\|_1 - \gamma^T y \\ & = -\max_{y} \gamma^T y - \lambda \|y\|_1 \\ & = 0 \quad \text{if} \ \|\gamma\|_\infty \leq \lambda, \\ & \text{or} = -\infty \quad \text{otherwise} \\ & \Downarrow \\ & \|\gamma\|_\infty \leq \lambda. \end{aligned} \tag{10}$$

When optimal (minimizing) conditions (see (8), (9), and (10)) are replaced in dual formulation (7), it becomes

$$\begin{aligned} \max_{\alpha\geq 0,\alpha\leq c\mathbf{1},\|\gamma\|_\infty\leq\lambda} \quad & \text{Dual}(\alpha, \gamma) \\ & = \frac{1}{2} \left(D^T\alpha - \gamma \right)^T Q^{-1} Q Q^{-1} \left(D^T\alpha - \gamma \right) \\ & \quad - \alpha^T D Q^{-1} \left(D^T\alpha - \gamma \right) \\ & \quad + \gamma^T Q^{-1} \left(D^T\alpha - \gamma \right) + \mathbf{1}^T\alpha. \end{aligned} \tag{11}$$

After negating (11) to turn it into minimization problem and after simplification of the expression, final problem formulation is

$$\begin{aligned} \min \quad & \frac{1}{2} \left(D^T\alpha - \gamma \right)^T Q^{-1} \left(D^T\alpha - \gamma \right) - \mathbf{1}^T\alpha \\ \text{s.t.} \quad & \mathbf{0} \leq \alpha \leq c\mathbf{1}, \\ & -\lambda\mathbf{1} \leq \gamma \leq \lambda\mathbf{1}. \end{aligned} \tag{12}$$

The original nonsmooth problem is turned into the smooth dual problem, which can be solved for its two sets of parameters α and γ. Since the strong duality holds, a solution to dual is a solution to the original problem, and optimal weight vector w^* can be retrieved after plugging the solution of dual, α^* and γ^*, into (8).

Similar dual formulation, just without the dummy variables y and associated multipliers γ, might be used for DSSL with the exact Hinge loss, instead of the originally proposed DSSL which uses Huber loss approximation [6].

2.3. Optimization Algorithm. The differentiable dual from (12) is, in fact, a quadratic optimization problem with box constraints:

$$\begin{aligned} \min_{x} \quad & \frac{1}{2} x^T H x + \mathbf{f}^T x \\ \text{s.t.} \quad & \mathbf{lb} \leq x \leq \mathbf{ub}, \end{aligned} \tag{13}$$

$$x = \begin{bmatrix} \alpha \\ \gamma \end{bmatrix},$$

$$\mathbf{f} = \begin{bmatrix} \mathbf{1}_{k\times 1} \\ \mathbf{0}_{d\times 1} \end{bmatrix},$$

$$\mathbf{lb} = \begin{bmatrix} \mathbf{0}_{k\times 1} \\ -\lambda \mathbf{1}_{d\times 1} \end{bmatrix}, \tag{14}$$

$$\mathbf{ub} = \begin{bmatrix} c\mathbf{1}_{k\times 1} \\ \lambda \mathbf{1}_{d\times 1} \end{bmatrix},$$

$$H = \begin{bmatrix} D & \mathbf{0}_{k\times d} \\ \mathbf{0}_{d\times d} & -I \end{bmatrix} \begin{bmatrix} Q^{-1} & Q^{-1} \\ Q^{-1} & Q^{-1} \end{bmatrix} \begin{bmatrix} D^T & \mathbf{0}_{d\times d} \\ \mathbf{0}_{d\times k} & -I \end{bmatrix}.$$

There are ready-to-use tools for solving the problem in (13), and we utilized the built-in Matlab "quadprog" solver, which is implemented as a projection method with the active set.

3. Results

3.1. Severity Score Characterization on Synthetic Data. For the initial assessment of the proposed framework, we have generated a synthetic example with properties that motivated the approach. If a large number of variables are measured, many are expected to be irrelevant for the assessment of severity.

We defined the severity score as a linear combination of intensities of the first 10 features after initiating a set of 100. In addition, we set the coefficients to have different magnitudes, as it is expected that contribution of different variables is of various levels (Figure 1(a)). The remaining ninety features do not affect severity score at all; they are irrelevant and only introduce uncertainty into the problem. For training purposes, values of all features are randomly sampled from a uniform distribution for 10 fictitious subjects with 10 different measurements each. Severity scores are associated based on a linear function with weights depicted in Figure 1. Comparison labels (pairs) were generated as all possible pairs in which the first element (sample) has substantially higher severity score as compared to the second element. This requirement of substantial gap in severity between pairs serves to mimic the case where a doctor could claim, with high confidence, that one patient is in more severe condition than another. Such generated training data was utilized to fit Sparse LDSS, (dense) DSSL, and DSSL model trained on the exact 10 features that are relevant, which we named Ideal DSSL in Table 1.

All models were tested on comparison pairs from an additional 50 test subjects with 10 measurements each. Testing data was generated by the same protocol as explained for training, except the threshold for the required difference of scores was set several times lower, in order to see how learned

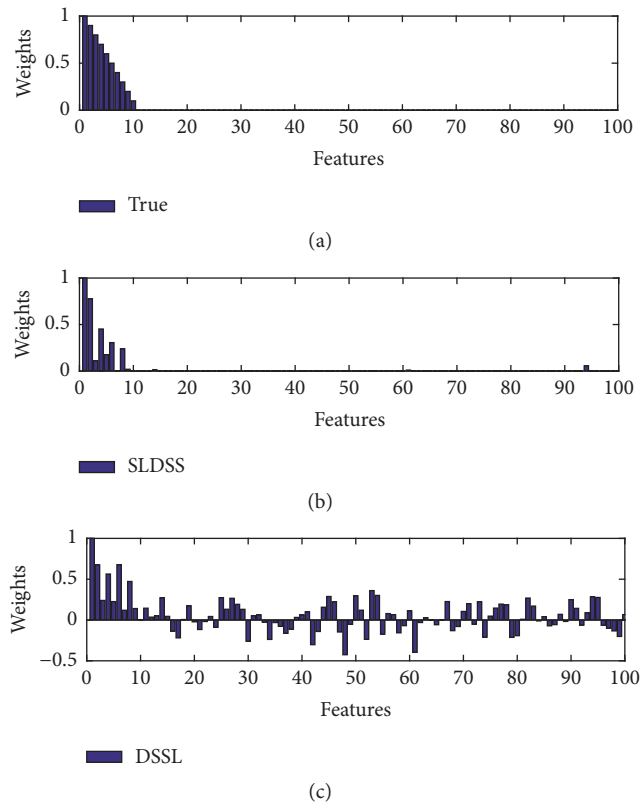

Figure 1: Synthetic example. Comparison of learned weight vectors (normalized) of sparse SLDSS method and dense DSSL method with the ground truth.

Table 1: Performance on synthetic data as measured by correctly ordered pairs, accuracy, and by aggregated error (magnitude of difference in wrongly ordered pairs), Hinge loss.

Approach	SLDSS	DSSL	Ideal DSSL
Accuracy	0.9397	0.8373	0.9558
Hinge loss	176.06	3110.20	180.65

functions generalize to more subtle differences between the cases.

$$\begin{aligned} \text{Accuracy} &= \frac{\#\ \text{Correctly Ranked}}{\#\ \text{Total Examples}} \\ &= 1 - \frac{\#\ \text{Incorrectly Ranked}}{\#\ \text{Total Examples}}. \end{aligned} \tag{15}$$

The predictive performance was measured as "Accuracy" (see (15)), that is, the fraction of the total examples that are correctly ordered, meaning that a linear function assigned a higher score to the first component of a pair. The results presented in Table 1 show that learning a dense weight vector impairs the predictive accuracy of the model, while learning a sparse vector, using the SLDSS, approaches the accuracy of the Ideal model, obtained by learning a disease severity score from relevant features known in advance. Figure 1 shows the

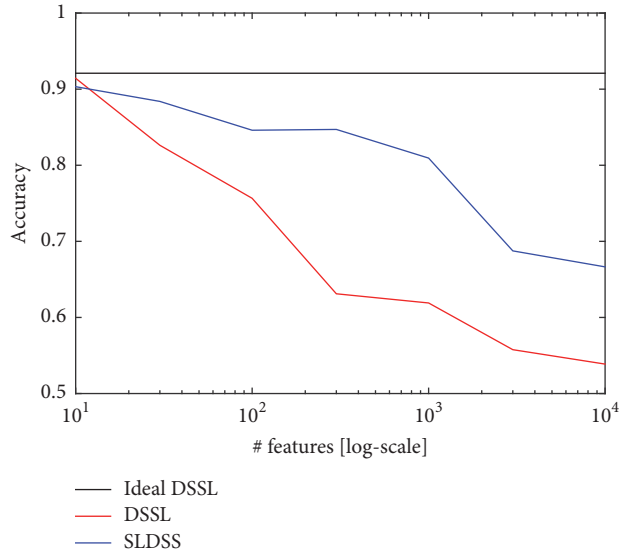

Figure 2: Synthetic experiment. Influence of the problem dimensionality (number of features) on the accuracy of ranking methods.

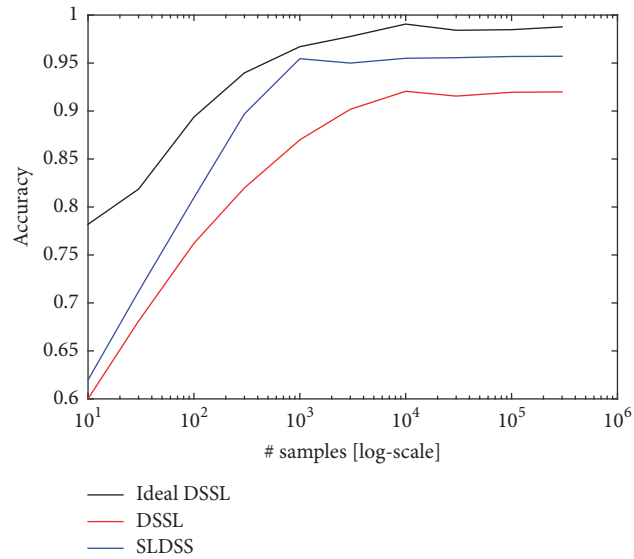

Figure 3: Synthetic experiment. Influence of the sample size (number of sample pairs) on the accuracy of ranking methods.

weights of learned severity functions, and it might be seen that reason for the reduced testing accuracy of the dense DSSL method (Figure 1(c)) is because it assigned nonzero weights to (by design) completely irrelevant features.

3.2. Feature Size Analysis. We have explored how the number of irrelevant features affects the model performance. This time we sampled 100 subjects (with 10 time-step samples each), with 10,000 features, where only the first 10 contribute to the true score. We varied the number of features from 10 (all features informative) up to 10,000 in exponentially progressive increments [10; 30; 100; 300; 1,000; 3,000; 10,000]. Results presented in Figure 2 show that when all features are informative (10 out of 10) DSSL is slightly better than SLDSS. However, as soon as irrelevant features are added, the SLDSS approach becomes more accurate than DSSL. As more irrelevant dimensions are added, both approaches' performance decreases, however SLDSS at a slower pace.

3.3. Sample Size Analysis. We also investigated how the number of training samples affects the predictive performance of the ranking approaches. We generated another synthetic set of 100 subjects (10 samples each). All samples had 100 features, where the first 10 were relevant for the ground truth score. From such generated examples, we constructed 357,355 comparison pairs for training. We varied the number of sample pairs, by randomly sampling from 10 up to 300,000 in exponentially progressive increments [10; 30; 100; 300; 1,000; 3,000; 10,000; 30,000; 100,000; 300,000]. From the results on holdout testing set, presented in Figure 3, it can be seen that accuracy increases with the number of training pairs and that SLDSS is always more accurate than DSSL. The Ideal DSSL, which is always trained only on the 10 relevant features, is consistently the most accurate.

3.4. Severity Score for Influenza A Virus. To further assess the proposed approach, we applied it to learning the severity of H3N2 influenza symptoms. The utilized dataset (http://people.ee.duke.edu/~lcarin/reproduce.html) contains temporally collected gene expression measurements of human subjects infected with H3N2 virus [16]. The samples were collected on multiple occasions (approximately every eight hours) during the period of one week after the virus was inoculated in subjects. Concurrently, the severity of their symptoms was tracked (approximately twice a day) and clinically assessed using the modified Jackson score [17]. When measurement time points were not perfectly aligned with severity score estimates, the temporally nearest estimate was associated with the gene expression vector. Having high dimensionality of the measurements (12,032 genes), temporally collected samples, and associated severity score estimates, this dataset was suited for testing the proposed severity score learning framework. In addition to direct assessments of severity scores, which could be used for regression, the data samples are also accompanied with class labels "symptomatic" and "asymptomatic" [18], based on the values of modified Jackson scores. Our comparison pairs generation process follows the guidelines proposed in [6]. Ideally, an expert would be presented with example pairs and would assess which one appears more intense (with respect to a property of interest), based on visual inspection, clinical report, or arbitrary convenient source. The alternative is to use an existing scoring system to generate comparison pairs, and for this application we utilized the Jackson score. We generated a third label type by extracting all possible pairs of samples where the first component is associated with a score that is substantially larger than the second. In our experiments, the "substantial" is defined by setting a threshold to 5 for training and 1 for testing.

Table 2: Performance on H3N2 influenza gene expression dataset as measured by the fraction of correctly ordered pairs (accuracy).

Approach	SLDSS	DSSL	LASSO	L_1 Logistic Regression
Accuracy	0.8097	0.7689	0.9490	0.7815

On the described dataset containing 267 samples (17 subjects with about 16 temporal samples each), we have compared the predictive performances of 4 methods:

(1) Sparse Learning of Disease Severity Score (SLDSS) from comparison pairs

(2) "Dense" Disease Severity Score Learning (DSSL) from comparison pairs

(3) LASSO regression on direct values of severity scores

(4) L_1-regularized Logistic Regression fitted on binary classification labels of symptom severity.

All enumerated methods result in a vector of feature weights that can be used as a scoring function. Except for the DSSL which results in a dense vector of weights, all other approaches typically only have a small number of nonzero weights, while all others are exactly equal to zero.

We compared the mentioned methods in a 10-fold cross-validation procedure (where all samples belonging to one subject are either all in training or all in testing folds) and the results are shown in Table 2.

In conducted experiments, the nonsparse method (DSSL) has the lowest accuracy, which provides evidence that sparse approaches were beneficial. LASSO was the most accurate, due to its direct access to the ground truth values (of the underlying scores), while other methods only had access to partial information. The Logistic Regression only had information if the score was larger than a certain threshold, while the DSSL and SLDSS only knew, for a list of pairs, which element in a given pair had a higher score. This, on the other hand, limits the application of LASSO to cases where scoring function already exists, thus reducing the necessity for learning it from the data. Among the approaches which learn from indirect information about underlying values of scores (comparison pairs and severity classes), our SLDSS is the most accurate.

3.5. Robustness of Selected Features. We were also interested in using SLDSS for feature selection to discover the most relevant variables for the condition. Therefore, we have performed additional analysis regarding the robustness (stability) of the selected features. Robustness of selected features is a very important aspect of the feature selection algorithms that was relatively neglected up until recently [19]. Various fields aim at finding the right subset of variables that would allow reliable prediction, and the more there are candidates to search from, the harder it is to find the right subset. Feature selection methods play a crucial role there, but when the dimensionality of data is much higher than the number of samples, the expectation of consistently finding high-quality solution decreases [20]. On the other side, L_1 regularized models have far fewer requirements for sample size as compared to rotation invariant models (L_2 regularized models, Support Vector Machines, Artificial Neural Networks, and DSSL, whose sample complexity grows at least linearly in the number of irrelevant features), as their sample size requirement grows logarithmically in the dimension of (irrelevant) features [21], so they are an attractive tool for such tasks.

Robustness is a metric that quantifies how different training sets affect the affinity of the algorithm towards the particular features and there are different measures proposed [22]. Here we used the common three:

(1) Pearson coefficient (see (16)), which measures the correlation between the weight vectors w and w' learned on different data (sub)sets and tells magnitude stability of the weights. In the case when the weight vector is used as a linear function, it also tells how stable the learned function is.

$$C_P\left(w, w'\right) = \frac{\sum_i \left(w_i - \mu_w\right)\left(w_i' - \mu_{w'}\right)}{\sqrt{\sum_i \left(w_i - \mu_w\right)^2 \sum_i \left(w_i' - \mu_{w'}\right)^2}}. \tag{16}$$

(2) Spearman rho metric (see (17)), which measures how well the orders (ranks) r and r' of weights' w and w' magnitudes are preserved between different training sets. It is important, for example, in the dense methods where features are selected as some top number of features according to the magnitude of weights.

$$C_S\left(r, r'\right) = \frac{\sum_i \left(r_i - \mu_r\right)\left(r_i' - \mu_{r'}\right)}{\sqrt{\sum_i \left(r_i - \mu_r\right)^2 \sum_i \left(r_i' - \mu_{r'}\right)^2}}. \tag{17}$$

(3) Jaccard index (see (18)), which measures the overlap between two discrete sets s and s' of nonzero features in w and w', normalized with their union ($|\cdot|$ is cardinality operator). Jaccard index is the most relevant measure (out of the three mentioned) regarding the stability of selected features, as studied frameworks select features in the form of a discrete set of nonzero features.

$$C_J\left(s, s'\right) = \frac{\left|s \cap s'\right|}{\left|s \cup s'\right|} = \frac{\left|s \cap s'\right|}{|s| + \left|s'\right| - \left|s \cap s'\right|}. \tag{18}$$

All four severity score learning methods are assessed for consistency/robustness based on each of the three stability measures (see (16)–(18)), through a 10-fold cross-validation procedure on H3N2 data. The sparsity level was tuned with free parameters (for sparse methods) so as to produce the average number (over tenfold) of nonzero features of about 100 out of 12,032 possible (SLDSS 97.1±16.7; LASSO 99.9±8.8; L_1LogReg 101.7 ± 22.7), with results presented in Figure 4 and summarized in Table 3. The dense method, DSSL, was compared to others, according to Jaccard index, by taking only the top 100 features according to the largest magnitudes

Table 3: Stability of selected feature subsets summarized as an average pairwise similarity over ten training folds.

Measure	SLDSS	DSSL	LASSO	L_1 Logistic Regression
Pearson coefficient	**0.8656**	0.7402	0.7362	0.5562
Spearman rank	**0.8163**	0.7204	0.5162	0.3988
Jaccard index	**0.6916**	0.2946	0.3595	0.2474

in each of the folds separately. The results show that here proposed SLDSS method is the most stable one according to each of the three measures. This means that it learns the most stable severity score function (according to Pearson correlation), as well as the most stable set of nonzero features (according to Jaccard index). This evidence is suggesting that SLDSS is finding the most reliable signal in the data, out of all the tested approaches. Nevertheless, there are no guarantees that the selected set of features is free of false positives, as previously it was theoretically concluded that LASSO-like approaches select a superset of the true features [23].

3.6. Gene Ontology Overrepresentation Analysis. To further check the appropriateness of SLDSS method as a biomarker discovery tool, we performed gene ontology overrepresentation analysis to assess the relevance of a set of features extracted from the influenza dataset. In the robustness analysis section, we found that more than two-thirds (0.6916) of the nonzero features are, on average, shared between the different folds of data. In fact, 50 genes were nonzero in all of the folds, so we took that set of genes and submitted it for overrepresentation analysis in the PANTHER [24] online tool.

We analyzed the list of 50 selected genes given in Table 4, against all the 12,032 genes in the dataset. Some of the 12,032 genes were duplicates, and some symbols were not recognized by the database (annotation version and release date: GO Ontology database, released 2016-03-25) resulting in the comparison of the 50 selected genes against the reference list of 10,792 genes using the PANTHER Overrepresentation Test (released 2016-03-21) with Bonferroni correction. Bonferroni correction [25] is a simple and common method for multiple testing correction of significance value indicators. It is well acknowledged that it might be substantially conservative, especially when multiple tests are not independent. In multiple gene ontology process testing, it might be extremely conservative because descendants of a process are completely dependent on their parents. Nevertheless, even after overly conservative adjustments, a number of processes are found statistically significantly overrepresented with the cutoff value of 0.05 for P value. Significantly overrepresented GO biological processes (listed in Table 5) are related almost exclusively to immune response and a reaction of the host body to the virus. This is consistent with the fact that the dataset is about the response to viral infection, suggesting that the discovered set of features is indeed relevant for the studied process.

Table 4: Genes selected by the Sparse Disease Severity Score Learning method, listed in alphabetical order.

Gene symbols
AIM2
ALDH1A1
ATF3
BLVRA
C3AR1
CASP5
CASP7
CCL2
CCL8
CDKN1C
CXCL10
EIF2AK2
EPB41L3
ETV7
GBP1
HERC5
IFI35
IFI44
IFI44L
IFI6
IFIH1
IFIT1
IFIT2
IFIT3
IL18RAP
ISG15
LAMP3
LAP3
LILRA5
MAFB
MS4A4A
MX1
MYOF
OAS1
OAS2
OAS3
OASL
RIN2
RSAD2
RTP4
S100A12
SERPING1
SIGLEC1
STAT1
TFEC
TLR7
TNFSF10
TYMP
XAF1
ZBP1

4. Conclusion

We assessed multiple approaches to learning the severity scores in high-dimensional applications. Our results point

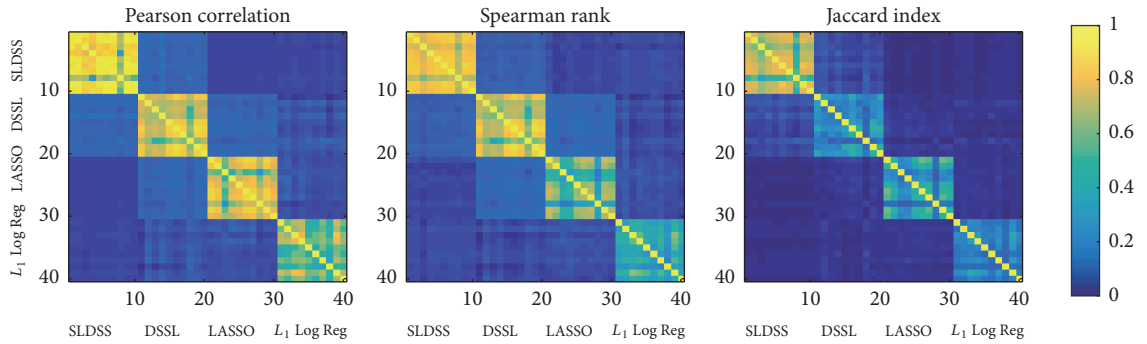

FIGURE 4: Similarity matrices. Similarity matrices between weight vectors learned over all 10 folds of data and all four methods. Warmer colors correspond to higher similarity (stability) and cooler tones to lower similarity. SLDSS (upper left squares) has the highest similarities among all methods.

TABLE 5: PANTHER overrepresentation analysis results. No.: number of associated genes; Exp.: expected number of genes by chance; Fold: number of times enriched.

GO biological process complete	GOID	No.	Exp.	Fold	*P* value
Defense response to virus	(GO:0051607)	18	.62	29.21	$4.00E-18$
Response to virus	(GO:0009615)	20	.96	20.75	$1.01E-17$
Type I interferon signaling pathway	(GO:0060337)	14	.28	49.54	$1.97E-16$
Cellular response to type I interferon	(GO:0071357)	14	.28	49.54	$1.97E-16$
Response to type I interferon	(GO:0034340)	14	.29	47.96	$3.08E-16$
Immune response	(GO:0006955)	32	5.42	5.91	$2.37E-15$
Immune system process	(GO:0002376)	35	8.20	4.27	$4.07E-13$
Innate immune response	(GO:0045087)	26	3.79	6.87	$1.07E-12$
Defense response	(GO:0006952)	31	6.12	5.07	$1.07E-12$
Defense response to other organisms	(GO:0098542)	19	1.73	10.97	$1.46E-11$
Immune effector process	(GO:0002252)	19	1.75	10.85	$1.76E-11$
Cytokine-mediated signaling pathway	(GO:0019221)	20	2.13	9.40	$3.84E-11$
Cellular response to cytokine stimulus	(GO:0071345)	21	2.72	7.72	$3.01E-10$
Response to cytokine	(GO:0034097)	22	3.16	6.96	$4.80E-10$
Response to other organisms	(GO:0051707)	22	3.23	6.81	$7.44E-10$
Response to external biotic stimulus	(GO:0043207)	22	3.23	6.81	$7.44E-10$
Response to biotic stimulus	(GO:0009607)	22	3.31	6.65	$1.21E-09$
Negative regulation of viral genome replication	(GO:0045071)	8	.19	41.11	$1.83E-07$
Negative regulation of viral process	(GO:0048525)	9	.36	24.90	$7.74E-07$
Regulation of viral genome replication	(GO:0045069)	8	.30	26.56	$5.56E-06$
Negative regulation of viral life cycle	(GO:1903901)	8	.35	23.02	$1.69E-05$
Response to stress	(GO:0006950)	34	14.20	2.40	$5.49E-05$
Negative regulation of multiorganism process	(GO:0043901)	9	.60	15.06	$6.01E-05$
Response to external stimulus	(GO:0009605)	26	8.48	3.06	$1.19E-04$
Cellular response to interferon-gamma	(GO:0071346)	8	.50	16.14	$2.59E-04$
Regulation of viral process	(GO:0050792)	9	.79	11.43	$6.26E-04$
Response to interferon-gamma	(GO:0034341)	8	.57	13.93	$7.95E-04$
Regulation of symbiosis	(GO:0043903)	9	.88	10.17	$1.66E-03$
Regulation of viral life cycle	(GO:1903900)	8	.74	10.86	$5.14E-03$
Interferon-gamma-mediated signaling pathway	(GO:0060333)	6	.31	19.33	$5.39E-03$
Response to stimulus	(GO:0050896)	43	26.53	1.62	$6.68E-03$
Regulation of defense response	(GO:0031347)	14	3.03	4.61	$8.01E-03$
Regulation of cytokine production	(GO:0001817)	12	2.19	5.48	$9.59E-03$
Cellular response to organic substance	(GO:0071310)	23	8.36	2.75	$1.01E-02$
Regulation of multiorganism process	(GO:0043900)	11	1.81	6.07	$1.07E-02$
Response to interferon-alpha	(GO:0035455)	4	.09	45.44	$1.55E-02$
Cellular response to chemical stimulus	(GO:0070887)	25	10.06	2.49	$1.75E-02$
Cell surface receptor signaling pathway	(GO:0007166)	23	9.04	2.54	$4.07E-02$
Multiorganism process	(GO:0051704)	22	8.40	2.62	$4.59E-02$

to the utility and maybe even necessity of reducing the dimensionality of the problem through sparse learning techniques. Combination of the advantages of existing solutions turned out to be beneficial for the performance of the formulation proposed in this study. The robustness of the learned disease severity score function, as well as features selected by our approach, compares very favorably to the alternatives. Conducted gene ontology overrepresentation analysis supports the relevance of the genes identified by the SLDSS approach. Additional studies are possible to further characterize selected genes and the processes they are involved in, in order to provide further insight into causal relations underlining the influenza infection. These are all mounting evidence that our approach could be used as a discovery tool for both disease severity scores and related informative variables, which could further motivate novel hypotheses. The method we proposed in this article is appropriate for learning severity score from a relatively small number of high-dimensional cases. More efficient optimization tools would be needed for applications where the number of cases is also large since in such applications a quadratic number of comparisons in the number of samples can be a challenge.

Abbreviations

SOFA:	Sequential Organ Failure Assessment
ICU:	Intensive Care Unit
APACHE:	Acute Physiology and Chronic Health Evaluation
SVM:	Support Vector Machine
LASSO:	Least Absolute Shrinkage and Selection Operator
SLDSS:	Sparse Linear Disease Severity Score
DSSL:	Disease Severity Score Learning.

Disclosure

The funding bodies had no role in the design, collection, analysis, or interpretation of this study, and any opinions, findings, conclusion, or recommendations expressed in this material are those of the authors and do not necessarily reflect the view of the DARPA, ARO, or the US Government.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this article.

Authors' Contributions

Ivan Stojkovic developed and implemented the computational methods and conducted the experiments, supervised by Zoran Obradovic. Ivan Stojkovic and Zoran Obradovic discussed and analyzed the results and wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgments

The authors wish to thank Aleksandar Obradovic for proofreading and editing the language of the manuscript. This material is based upon work partially supported by the Defense Advanced Research Projects Agency (DARPA) and the Army Research Office (ARO) under Contract no. W911NF-16-C-0050 and partially supported by DARPA Grant no. 66001-11-1-4183 negotiated by SSC Pacific grant.

References

[1] J.-L. Vincent, R. Moreno, J. Takala et al., "The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure," *Intensive Care Medicine*, vol. 22, no. 7, pp. 707–710, 1996.

[2] W. A. Knaus, E. A. Draper, D. P. Wagner, and J. E. Zimmerman, "APACHE II: a severity of disease classification system," *Critical Care Medicine*, vol. 13, no. 10, pp. 818–829, 1985.

[3] S. Yang, L. Shapiro, M. Cunningham et al., "Skull retrieval for craniosynostosis using sparse logistic regression models," in *Medical Content-Based Retrieval for Clinical Decision Support*, pp. 33–44, Springer, 2012.

[4] J. Zhou, J. Liu, V. A. Narayan, and J. Ye, "Modeling disease progression via fused sparse group lasso," in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '12)*, pp. 1095–1103, August 2012.

[5] K. Dyagilev and S. Saria, "Learning severity score for sepsis: a novel approach based on clinical comparisons," in *Proceedings of the in. AMIA Annual Symposium Proceedings*, pp. 1890–1898, 2015.

[6] K. Dyagilev and S. Saria, "Learning (predictive) risk scores in the presence of censoring due to interventions," *Machine Learning*, vol. 102, no. 3, pp. 323–348, 2016.

[7] T. Joachims, "Optimizing search engines using clickthrough data," in *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 133–142, July 2002.

[8] I. M. Johnstone and D. M. Titterington, "Statistical challenges of high-dimensional data," *Philosophical Transactions of the Royal Society A: Mathematical, Physical & Engineering Sciences*, vol. 367, no. 1906, pp. 4237–4253, 2009.

[9] R. Tibshirani, "Regression shrinkage and selection via the lasso: a retrospective," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 73, no. 3, pp. 273–282, 2011.

[10] D. Ghosh and A. M. Chinnaiyan, "Classification and selection of biomarkers in genomic data using LASSO," *Journal of Biomedicine and Biotechnology*, vol. 2005, no. 2, pp. 147–154, 2005.

[11] I. Stojkovic, M. F. Ghalwash, and Z. Obradovic, "Ranking based multitask learning of scoring functions," in *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, 2017.

[12] A. J. Atkinson Jr., W. A. Colburn, V. G. DeGruttola et al., "Biomarkers and surrogate endpoints: preferred definitions and conceptual framework," *Clinical Pharmacology & Therapeutics*, vol. 69, no. 3, pp. 89–95, 2001.

[13] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005.

[14] C. De Mol, E. De Vito, and L. Rosasco, "Elastic-net regularization in learning theory," *Journal of Complexity*, vol. 25, no. 2, pp. 201–230, 2009.

[15] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, Cambridge, UK, 2004.

[16] A. K. Zaas, M. Chen, J. Varkey et al., "Gene expression signatures diagnose influenza and other symptomatic respiratory viral infections in humans," *Cell Host & Microbe*, vol. 6, no. 3, pp. 207–217, 2009.

[17] G. G. Jackson, H. F. Dowling, I. G. Spiesman, and A. V. Boand, "Transmission of the common cold to volunteers under controlled conditions: I. The common cold as a clinical entity," *A.M.A Archives of Internal Medicine*, vol. 101, no. 2, pp. 267–278, 1958.

[18] C. W. Woods, M. T. McClain, and M. Chen, "A host transcriptional signature for presymptomatic detection of infection in humans exposed to influenza H1N1 or H3N2," *PLoS ONE*, vol. 8, no. 1, Article ID e52198, 2013.

[19] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007.

[20] C. Sima and E. R. Dougherty, "What should be expected from feature selection in small-sample settings," *Bioinformatics*, vol. 22, no. 19, pp. 2430–2436, 2006.

[21] A. Y. Ng, "Feature selection, L 1 vs. L 2 regularization, and rotational invariance," in *Proceedings of The Twenty-First International Conference on Machine Learning*, Banff, Alberta, Canada, July 2004.

[22] A. Kalousis, J. Prados, and M. Hilario, "Stability of feature selection algorithms: a study on high-dimensional spaces," *Knowledge and Information Systems*, vol. 12, no. 1, pp. 95–116, 2007.

[23] P. Bühlmann and S. van de Geer, *Statistics for High-dimensional Data: Methods, Theory and Applications*, Springer Series in Statistics, Springer, New York, NY, USA, 2011.

[24] H. Mi, S. Poudel, A. Muruganujan, J. T. Casagrande, and P. D. Thomas, "PANTHER version 10: expanded protein families and functions, and analysis tools," *Nucleic Acids Research*, vol. 44, no. D1, pp. D336–D342, 2016.

[25] W. Haynes, "Bonferroni Correction," in *Encyclopedia of Systems Biology*, W. Dubitzky, O. Wolkenhauer, K.-H. Cho, and H. Yokota, Eds., pp. 154-154, Springer, New York, NY, USA, 2013.