Combining Machine Learning and Statistical Modeling to Identify Risk Factors of Hospital Mortality and Directionality for Patients with Acute Respiratory Distress Syndrome (ARDS)

Meng Zhang, Michael Qiu, Molly Stuart, Jamie S. Hirsch, Negin Hajizadeh

INTRODUCTION:

Identifying factors associated with patient outcomes in clinical studies and in retrospective data analysis informs further targeted clinical studies and basic science research. Classical regression-based statistical models such as linear regression, logistic regression, etc. were typically applied to establish and quantify these relationships between risk factors and patient outcomes through statistical inference.[1] However, these analyses were usually performed with hypothesis-constrained datasets where candidate variables were limited by research questions of interest (and therefore selection of variables to measure in studies) which in turn were informed by clinical observations or prior studies. [2] Therefore, factors with unknown/unanticipated associations will be missed using this approach, and may never be discovered.

An Electronic Health Record (EHR) is a digital version of a patient's paper chart. EHRs contain comprehensive patient data including medical history, diagnoses, medications, treatment plans, immunization dates, allergies, radiology images, and laboratory and test results which provide a more complete picture of patients' health status. [3] Large volumes of data are readily available within EHRs and can be analyzed in a non-hypothesis constrained way. This data-driven approach may discover novel risk factors of patient outcomes, which may be superior to hypothesis-driven approaches. However, it is challenging to analyze the large volumes of data using classical statistical models, especially when the sample size is relatively small compared to the number of predictors. In classical regression analysis, the common approach to statistical model building is to minimize the number of variables until the most parsimonious model that describes the data is found, which also results in numerical stability and generalizability of the results. [4][5] With a big number of predictor variables, it is hard to weed out irrelevant and redundant variables to identify predictors truly likely to be contributing to the outcome. [6]

One logical first approach to deal with the big number of predictor variables is to start with screening tests of these variables, retaining those that meet a pre-specified criterion. One of the most commonly used methods is to do univariate analyses such as t-test or Chi-squared test or bivariate regression models to look for independent associations (i.e. without consideration of other covariates) with the outcome. [4] Variables that meet a criterion, such as p-value less than 0.05, are retained in the data for further multivariate regression analysis. Multivariate regression models allow for the estimation of marginal effects of multiple variables at the same time, after controlling for the effects of other variables on the outcomes.[1] In the meantime, variables that do not meet the criterion are discarded. However, when the predictor variables have large interaction effects but small marginal effects, they might not be found to be statistically significant in the univariate analysis and are likely to be missed for further analysis. [7] This might be remedied by 'forcing' variables into the multivariate models usually based on prior studies and/or clinical suspicion. [4] But when there is no such information available for these variables, they are unlikely to ever be evaluated as risk factors. On the other hand, when the sample size is large, many variables may be found to be statistically significant in the univariate analysis, even though the strength of their association with patient outcomes may not be clinically significant. [8] These variables are usually nuisance variables which may create noise and distort the strength of association between true risk factors and patient outcomes in the final multivariate regression models. [9]

Another method of screening for candidate variables for the final model is to use stepwise regression to search a large space of possible models for the best subset of variables through exhaustive searches of all possible combinations of these variables. [10] However, this method is computationally intensive, in particular when analyzing large datasets with thousands of variables, such as is now possible with electronic health record data.

Machine learning techniques such as Random Forest, use computation methods to "learn" information directly from data and mainly focus on prediction of a certain outcome. [11] In clinical studies, they are applied to predict patient outcomes with large datasets including EHR data. [12] Unlike classic regression analysis, it is usually not straightforward to summarize the relationship between the predictors and outcome into a single parameter using these techniques. Therefore, they are often referred to as "a black box" "due to lack of interpretability of the relationship between the risk factors and patient outcomes. [13]However, many machine-learning methods summarize the impact of individual variables into metrics referred to as variable importance, taking into account variable interactions without model specification required. [14] These techniques include Random Forest, [15] Support Vector Machine, [16] Gradient Boosting, [17] Lasso and Ridge regression. [18] The measure of variable importance can be ranked to indicate how important they are to the fitted model. Variables with high importance are potential drivers of the outcome and their values have a likely real impact on the outcome values. [15]

In the current study, we aimed to leverage these machine learning techniques of variable ranking, applied to a large EHR dataset, to identify variables most likely to be associated with patient outcomes for further multivariate analysis. We sought to explore: 1) whether there are differences in the candidate variables identified for multivariable analysis using machine learning methods, versus classic univariate analysis methods, versus a combination of machine learning methods and classic univariate analysis (i.e. application of classic univariate analysis to machine learning method identified top ranked variables) ; 2) After variable selection among the candidate variables identified using the three approaches above, whether there are differences in the variables included in the final multivariate models and whether the performances of these final models are different. Different machine learning techniques use different algorithms to make predictions. Accordingly, the calculation of variable importance in predictions of outcomes varies by different techniques. Different rank orderings should be expected among these techniques. Furthermore, different machine learning techniques sometimes result in similar performances. Therefore, in the initial variable screening step, instead of relying on variable importance ranks based on only one of the machine-learning techniques, we proposed to retain the predictor variables ranked high on average across these techniques, which is expected to identify potential predictors of the outcome in a more comprehensive way.

We used an EHR dataset for patients with severe Acute Respiratory Distress Syndrome (ARDS) to illustrate our proposed methodology and to explore our questions about differential outcomes. Hospital mortality for patients with severe ARDS remains high and new knowledge about unknown risk factors associated with hospital mortality could inform future targeted basic science and clinical studies.

METHODS:

Cohort and data structuring:

We retrospectively identified 246 patients with severe ARDS during the flu season (October to April) between 2016 and 2018 from EHR system across the multi-hospital Northwell Health system.

The criteria of diagnosis of severe ARDS of intubated patients included: age ≥ 18; diagnosis of severe acute respiratory failure requiring invasive mechanical ventilation via endotracheal tube or tracheostomy (PEEP ≥ 5 cm H2O); PaO2/FiO2 ratio (P : F ratio) ≥ 150; current or planned admission to an ICU; bilateral opacities on chest radiograph or computed tomography scan not fully explained by effusions, lung collapse or nodules; and respiratory failure not fully explained by cardiac failure or fluid overload based on the imaging studies completed close to the time when patient met P: F ratio. These criteria were manually screened by two physicians. The research project was approved by the Northwell Health institutional IRB.

We used patient information during the first 24 hours from severe ARDS diagnosis for this illustrative analysis, which included 107 baseline variables. The baseline patient information included demographics, comorbidities, ARDS risk factors and number of days between hospital admission and severe ARDS diagnosis. Other patient information included laboratory tests, medications, ventilation modes and adjuvant therapies. For lab tests, since multiple tests were usually performed, we calculated the median, minimum and maximum of the recorded values. For medications, we chose the highest dosage on a given day to represent the daily value. The primary response variable was hospital mortality (whether a patient died during the current hospitalization), encoded in a binary variable (1=Yes, 0=No). Regarding to the missing values, a missing category was created for categorical variables with missing data, and median of the available data was used to impute the missing values of continuous variables.

Analysis:

Three approaches to find the candidate variables for the final multivariate model were explored in our study:

1) Classic univariate analysis candidate variable selection

Classic univariate (bivariate logistic regression) analysis was performed among all the 107 variables in the data. Variables that were found statistically significant (p-value<0.05) in the univariate analysis were chosen as the candidate variables in the final multivariate logistic regression model.

2) Machine learning methods candidate variable selection

We used machine learning techniques to preliminarily screen candidate variables associated with hospital mortality (whether a patient died during the current hospitalization), encoded in a binary variable (1=Yes, 0=No)) by ranking variable importance among all the 107 variables. These techniques included random forests (RF), support vector machine (SVM), gradient boosting (GB), Lasso regression and Ridge regression.

For each of the machine learning techniques, the data was randomly split into a training set (80%) and a test set (20%). Within the training set, the data was split randomly into a sub-training set (70%) and a validation set (10%) for five times. Tuning parameters that impact the overall complexity of the final model and the final bias-variance trade-off were optimized by maximizing the average validation AUC (area under the Receiver Operating Curve) across the five validation sets. After the model was trained using the training set, it was applied to the test set. The test AUC was calculated with a 95% confidence interval using the DeLong method. The above process was

iterated for one hundred times within each machine learning technique. And, accordingly, the variable importance ranking was based on the model with the highest test AUC within each technique. We chose to calculate the final ranking of importance of these variables using the average rank across these five machine learning techniques.

Top ranked variables across model techniques were retained as the candidate variables in the subsequent multivariate analysis using classic logistic regression models, to identify risk factors highly associated with the outcome variable, and to determine whether the associations were protective or harmful. Given that there is no established cutoff for ranks of the variables to be included in multivariate models, we used the top 15% (17 variables), 20% (22 variables) and 25% (27 variables) separately.

3) Combination of machine learning methods and classic univariate analysis for candidate variable selection

Classic univariate (bivariate logistic regression) analysis was performed among all the top-ranked variables: 17 variables if ranked top 15%, 22 variables if ranked top 20%, and 27 variables if ranked top 25%. Variables that were found statistically significant (p-value<0.05) in the univariate analysis were chosen as the candidate variables in the final multivariate logistic regression model.

In all these three approaches, the final model for the multivariate analysis was constructed using backward selection among the candidate variables.

RESULTS:

Overall, among 246 patients identified with severe ARDS, 150 (60.98%) died during their hospitalization. 107 predictor variables were available for the analysis. Patient characteristics including demographics, comorbidities, ARDS risk factors and total SOFA scores of the study samples were are presented in Table 1. Among these variables, the variables that were associated with increased risk of in-hospital death included: age, and total SOFA score at baseline, and race; the variables that were associated with decreased risk of in-hospital death included: having diagnosis of pancreatitis and drug overdose. Specifically, compared to the patients who were discharged alive, patients who died in the hospital were older (65.58 ±16.85 vs. 53.85 ±17.82, p=0.0001), more likely not to have race reported (20.0% vs. 8.33%, p=0.02), had relatively higher total SOFA score at baseline (11.80 ±3.75 vs. 9.02 ±3.11, p=0.0001), less likely to have pancreatitis (2.67% vs. 10.42%) and less likely to have drug overdose (2.0% vs. 11.46%, p=0.002).

The five machine learning techniques performed similarly (Table 2), among which gradient boosting achieved highest test AUC of 0.84 (0.73, 0.95), followed by support vector machine and ridge regression, both with an AUC of 0.83 (0.72, 0.94), random forest with an AUC of 0.81 (0.70, 0.93), and Lasso regression 0.79 (0.66, 0.92). The variable importance of each of the 107 variables was ranked within each machine learning model and then averaged across the machine learning models.

The candidate variables primarily screened for the final multivariate model through the three approaches are listed in table 3. These variables are listed by rank based on the average ranking of variable importance across the five machine-learning techniques. There are a total of 33 variables listed in table 3. These variables include all the 26 variables among the 107 variables that were

found to be statistically significant in the univariate analysis, and all the 27 top-ranked variables (top 25%) based on average ranking of variable importance. There is a big overlap between these two lists of variables. Specifically, among the 26 variables that were found to be statistically significant in the univariate analysis, 22 variables (70%) ranked high (up to top 25%); and among the 27 top-ranked variables, 20 variables (74%) were also found to be statistically significant.

Table 4 shows the three final multivariate models constructed using the three approaches. The final models using approach #1 (classic) and #3 (combination) included exactly the same variables, while the final model using approach #2 (machine learning) included an additional variable (1st PEEP Value on Day 1) that was not captured by approach #1 and #3. This variable ranked high with machine learning methods (rank 11^{th}, among top 15%) yet was not statistically significant in the univariate analysis, but it was found to be statistically significant in the multivariate analysis after controlling for the other variables. The final model using approach #2 performed marginally better than the final model using approach #1 and #3 with C-statistic of 0.84 compared to 0.83.

DISCUSSION:

Machine learning techniques are typically applied in the prediction of the clinical outcomes with the capacity to analyze large datasets. However, due to the lack of transparency regarding to the mechanism of the risk factors identification, these techniques have limitations when quantifying the relationship between risk factors and clinical outcomes. [13] Traditional statistical modelling can help to further explain these associations. Our study leveraged these techniques to narrow candidate variables from a large dataset for further multivariate analysis, and compared the results with classic univariate analysis identification to see 1) if the variables ranked highly (up to 25%) could capture majority of variables that were found to be statistically significant in the classic univariate analysis, and further 2) if these variables could also capture variables that were not statistically significant in the classic univariate analysis, while found to be statistically significant in the multivariate analysis, so that unknown risk factors that might be missed through classic approach could be discovered.

Our findings using a sample dataset of severe ARDS patients demonstrated that in terms of the candidate variables for the final multivariate analysis, there was a big overlap between the list of variables that were found statistically significant in the classic univariate analysis (approach #1) and the list of variables that were ranked high (up to 25%) through machine learning method (approach #2). In other words, the machine learning method (approach #2) has the ability to capture majority of variables that were identified through classic univariate analysis (approach #1). In the meantime, although the final multivariable models constructed based on the three approaches were very similar, it was shown that the final model based on approach #2 (machine learning) captured one additional statistically significant variable that was not present in approach #1 (classic) and approach #3 (combination). This variable ranked high through machine learning variable ranking, while was found to be not statistically significant in the univariate analysis, so that it was missed by approach #1 and approach #3. This could be due to the fact that when the predictor variables have large interaction effects but small marginal effects, they might not be found to be statistically significant in the univariate analysis and are likely to be missed for further analysis. [16] Furthermore, the addition of this machine learning identified candidate variable in approach #2, marginally

improved the performance of the final model in approach #1 and approach #3 with C-statistic 0.84 vs. 0.83.

Overall, our findings showed that there are potential benefits of using average variable importance ranking in machine learning techniques to preliminarily screen candidate variables for further multivariate analysis in traditional statistical modelling. This approach may help identify variables that would have been missed/never have been discovered in the classic univariate analysis for further risk factor identification in the multivariate analysis. Also it might improve the performance of the final multivariable model compared to the approaches that are based on the classic univariate analysis.

However, the findings of this study was were only based on one illustrative data for exploratory purposes. Formal simulation study needs to be performed to compare these three approaches and see whether the machine learning candidate variable selection approach are superior to the other two approaches. Through the simulation study, we would like to see 1) what is the optimal cutoff of the ranks of the variables i.e. the percentage of the top ranked variable to be selected as the candidate variables for further multivariate analysis. This cutoff should capture majority of the variables that are found to be statistically significant in the classic univariate analysis, while retaining minimum numbers of variables for further multivariate analysis; 2) whether this approach can constantly identify variables that would have been missed/never have been discovered in the classic univariate analysis while turning out to be in the final multivariate model; 3) whether the final model constructed through this approach performs better or at least equally better compared to the other two approaches; 3) if the final model will perform better using approach #2 if different weights are given on the ranks of the variables among different machine learning techniques according to the performance for each individual technique, and how to define these weights.

Table 1: Characteristics of study sample

Variable	Discharged Alive (N=96)	Death at hospital (N=150)	Combined (N=246)	p-value
Age, (mean ± SD)	53.85 ±17.82	65.58 ±16.85	61.0 ±18.13	**0.0001**
Sex (Female), n (%)	42 (43.75%)	61 (40.67%)	103 (41.87%)	0.63
Race, n (%)				
American Indian or Alaskan Native	1 (1.04%)	0 (0%)	1 (0.41%)	0.39
Asian	7 (7.29%)	14 (9.33%)	21 (8.54%)	0.58
Black or African American	19 (19.79%)	22 (14.67%)	41 (16.67%)	0.29
Native Hawaiian or other Pacific Islander	0 (0%)	1 (0.67%)	1 (0.41%)	1.0
White	61 (63.54%)	83 (55.3%)	144 (58.54%)	0.20
Not reported	8 (8.33%)	30 (20.0%)	38 (15.45%)	**0.02**
Ethnicity, n (%)				
Hispanic	13 (13.54%)	12 (8.0%)	25 (10.16%)	0.16
Not Hispanic	73 (76.04%)	125 (83.33%)	198 (80.49%)	
Not reported	10 (10.42%)	13 (8.67%)	23 (9.35%)	0.65
Comorbidities, n (%)				
Cirrhosis	5 (5.21%)	9 (6.0%)	14 (5.69%)	0.79
Hepatic Failure	1 (1.04%)	5 (3.33%)	6 (2.44%)	0.41
End Stage Renal Disease requiring Hemodialysis	3 (3.13%)	14 (9.33%)	17 (6.91%)	0.06
Metastatic Carcinoma	1 (1.04%)	9 (6.0%)	10 (4.07%)	0.09
Lymphoma	3 (3.13%)	1 (0.67%)	4 (1.63%)	0.30
Leukemia	4 (4.17%)	9 (6.0%)	13 (5.28%)	0.53
Myeloma	0 (0%)	5 (3.33%)	5 (2.03%)	0.20
AIDS	0 (0%)	1 (0.67%)	1 (0.41%)	1.0
Immunosuppression	10 (10.42%)	13 (8.67%)	23 (9.35%)	0.65
Chronic lung disease	17 (17.71%)	29 (19.33%)	46 (18.70%)	0.75
Diabetes Mellitus	27 (28.13%)	30 (20.0%)	57 (23.17%)	0.14
Chronic Heart Failure	8 (8.33%)	15 (10.0%)	23 (9.35%)	0.66
Bone Marrow Transplant	0 (0%)	3 (2.0%)	3 (1.22%)	0.28

ARDS risk factors, n (%)				
Sepsis	66 (68.75%)	93 (62.0%)	159 (64.63%)	0.28
Pneumonia	68 (70.83%)	95 (63.33%)	163 (66.26%)	0.22
Aspiration	23 (23.96%)	36 (24.0%)	59 (23.98%)	0.99
Smoke Inhalation Injury			0	N.A.
Trauma	1 (1.04%)	2 (1.33%)	3 (1.22%)	1.0
Near drowning			0	N.A.
Pancreatitis	10 (10.42%)	4 (2.67%)	14 (5.69%)	**0.01**
Burn			0	N.A.
Shock	26 (27.08%)	58 (38.67%)	84 (34.15%)	0.06
Drug Overdose	11 (11.46%)	3 (2.0%)	14 (5.69%)	**0.002**
Blood Product Transfusion	17 (17.71%)	34 (22.67%)	51 (20.73%)	0.34
Other ARDS Risk Factors	6 (6.25%)	12 (8.0%)	18 (7.32%)	0.61
Total Sofa Scores, (mean ± SD)	9.02 ±3.11	11.80 ±3.75	10.72 ±3.76	**0.0001**

Table 2: Summary of model performance of each machine learning techniques (approach #2)

Machine Learning Techniques	Test AUC (95% CI)	Accuracy	Sensitivity (95% CI)	Specificity (95% CI)	PPV (95% CI)	NPV (95% CI)
Random Forest	0.81 (0.70, 0.93)	0.74	0.73 (0.54, 0.87)	0.76 (0.50, 0.93)	0.86 (0.67, 0.96)	0.59 (0.36, 0.79)
Support Vector Machine	0.83 (0.72, 0.94)	0.74	0.73 (0.54, 0.87)	0.76 (0.50, 0.93)	0.86 (0.67, 0.96)	0.59 (0.36, 0.79)
Gradiant Boosting	0.84 (0.73, 0.95)	0.74	0.74 (0.55, 0.88)	0.74 (0.50, 0.91)	0.82 (0.63, 0.94)	0.64 (0.41, 0.83)
Lasso Regression	0.79 (0.66, 0.92)	0.72	0.72 (0.53, 0.86)	0.72 (0.47, 0.90)	0.82 (0.63, 0.94)	0.59 (0.36, 0.79)
Ridge Regression	0.83 (0.72, 0.94)	0.74	0.76 (0.56, 0.90)	0.71 (0.48, 0.89)	0.79 (0.59, 0.92)	0.68 (0.45, 0.86)

Table 3: Candidate Variables primarily screened for the final multivariate model through three approaches

Variables as candidate variables in the final model	Ranks based on average ranking of variable importance across five machine-learning techniques	Approach #1: Including variables statistically significant in the univariate analysis among all the variables in the data	Approach #2: Including all the top ranked (15%, N=17; 20%, N=22; 25%, N=27) variables through average variable importance ranking	Approach #3: Including variables statistically significant in the univariate analysis among top ranked (25%) variables through average variable importance ranking
age	1	✓	✓	✓
Total SOFA Score	2	✓	✓	✓
Lowest Platelet Count (Day 1)	3	✓	✓	✓
Minimum P:F Ratio (Day 1)	4	✓	✓	✓
Median Peak Pressure (Day 1)	5	✓	✓	✓
Number of days between hospital admission and ARDS diagnosis	6		✓	
Minimum Peak Pressure (Day 1)	7	✓	✓	✓
Median Total Respiratory Rate (Day 1)	8	✓	✓	✓
Highest Creatinine (Day 1)	9	✓	✓	✓
Median P:F Ratio (Day 1)	10	✓	✓	✓
1st PEEP Value (Day 1)	11		✓	
Race Not Reported	12	✓	✓	✓
Median Tidal volume (cc) per kilogram of ideal	13		✓	

body weight (IBW)				
Maximum Total Respiratory Rate (Day 1)	14	✓	✓	✓
Maximum P:F Ratio (Day 1)	15	✓	✓	✓
Pancreatitis	16	✓	✓	✓
Maximum Peak Pressure (Day 1)	17	✓	✓	✓
Minimum Set Respiratory Rate (Day 1)	18			
Median Total Minute Ventilation (Day 1)	19	✓	✓	✓
Drug Overdose	20	✓	✓	✓
Received Norepinephrine (Day 1)	21	✓	✓	✓
Maximum Total Minute Ventilation (Day 1)	22	✓	✓	✓
Sepsis	23		✓	
Influenza test result (negative)	24		✓	
Lowest GCS Score (Day 1)	25	✓	✓	✓
Received HFNC prior to intubation	26	✓	✓	✓
Race: White	27		✓	
Mean Arterial Pressure (Day 1)	28	✓		
Received Vasopressors (Day 1)	30	✓		
Median Mean Airway Pressure (Day 1)	31	✓		

Minimum Total Respiratory Rate (Day 1)	34	✓		
Received Epinephrine (Day 1)	56	✓		
Maximum Mean Airway Pressure (Day 1)	67	✓		

Table 4: Final multivariate logistic models based on three variable screening approaches

Variable	Multivariate Analysis: Classic Logistic Regression Outcome: Hospital Mortality		
	Approach #1: Adjusted Odds Ratio	Approach #2 Adjusted Odds Ratio	Approach #3: Adjusted Odds Ratio
age	**1.042 (1.024, 1.061)**	1.043 (1.023, 1.063)	**1.042 (1.024, 1.061)**
Total SOFA Score	**1.265 (1.148, 1.393)**	1.307 (1.178, 1.449)	**1.265 (1.148, 1.393)**
Median Peak Pressure (Day 1)	1.079 (1.026, 1.135)	1.115 (1.053, 1.180)	1.079 (1.026, 1.135)
Median P:F Ratio (Day 1)	0.992 (0.986, 0.998)	**0.992 (0.986, 0.998)**	0.992 (0.986, 0.998)
1st PEEP Value (Day 1)		**0.809 (0.713, 0.918)**	
C statistic	0.83	0.84	0.83

REFERENCES:

1. Alexopoulos EC. Introduction to multivariate regression analysis. Hippokratia. 2010; 14 (Suppl 1):23–28.
2. Gliklich RE , Dreyer NA , Leavy MB . Registries for evaluating patient outcomes: a user's guide. 2014
3. Gunter TD, Terry NP. "The emergence of national electronic health record architectures in the United States and Australia: models, costs, and questions". Journal of Medical Internet Research 2015; 7 (1)
4. Bursac Z, Gauss CH, Williams DK, et al. Purposeful selection of variables in logistic regression. Source Code Biol Med. 2008; 3:17.
5. Shi, L.; Westerhuis, J.A.; Rosen, J.; Landberg, R.; Brunius, C. Variable selection and validation in multivariate modelling. Bioinformatics 2019; 35, 972–980
6. Galit Shmueli, Nitin R. Patel, Peter Bruce Wiley; Data Mining for Business Analytics: Concepts, Techniques, and Applications. Wiley, 2 edition 2010
7. Kathryn L Lunetta, L Brooke Hayward, Jonathan Segal, Paul Van Eerdewegh. "Screening large-scale association study data: exploiting interactions using random forests", BMC Genet. 2004; 5: 32
8. A.C. Skelly. "Probability, proof, and clinical significance". Evid Based Spine Care J, 2 (2011), pp. 9-11
9. Basu, D. "On the Elimination of Nuisance Parameters," Journal of the American Statistical Association 1977, vol. 77, pp. 355–366
10. Zhang Z. Variable selection with stepwise and best subset approaches. Ann Transl Med. 2016;4:136
11. Dey, A. Machine Learning Algorithms: A. Review. Int. J. Comput. Sci. Inf. Technol. 2016, 7, 1174–1179.
12. Rose, S. Machine Learning for Prediction in Electronic Health Data. JAMA Netw 2018. Open 1, e181404
13. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat. Mach. 2019, 1, 206–215
14. Alexander Zien, Nicole Krämer, Sören Sonnenburg, and Gunnar Rätsch. 2009. The feature importance ranking measure. In Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, 694–709
15. Louppe G, Wehenkel L, Sutera A, Geurts P (2013) Understanding variable importances in forests of randomized trees. In: Burges CJC, Bottou L, Welling M, Ghahramani Z, Weinberger KQ (eds) Advances in neural information processing systems, pp 431–439
16. Blanco R, et al. Gene selection for cancer classification using wrapper approaches, Int. J. Pattern Recognit. Artif. Intell. , 2004, vol. 18, pp. 1373-1390
17. F. Pan, T. Converse, D. Ahn, F. Salvetti, G. Donato, "Feature selection for ranking using boosted trees", Proc. 18th ACM Conf. Inf. Knowl. Manag. 2009, pp. 2025-2028
18. V. Fonti, E. Belitser, Feature selection using lasso, 2017.

19. Ho, Tin Kam (1995). Random Decision Forests (PDF). Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14–16 August 1995. pp. 278–282
20. Ho TK (1998). "The Random Subspace Method for Constructing Decision Forests" (PDF). IEEE Transactions on Pattern Analysis and Machine Intelligence. 20 (8): 832–844.
21. Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome (2008). The Elements of Statistical Learning (2nd ed.). Springer. ISBN 0-387-95284-5
22. Kawakubo H, Yoshida H (2012) Rapid feature selection based on random forest for high dimensional data. Expert Syst Appl 40:6241–6252
23. R. Berwick, "An Idiot's Guide to Support Vector Machines (SVMs)". Massachusetts Institute of Technology, Cambridge, MA, 2003
24. Friedman, J. (2001). Greedy function approximation: A gradient boosting machine. Ann. Statist. 29 1189--1
25. Natekin A., Knoll A., “Gradient boosting machines, a tutorial ”, in Frontier NeuroRobotics, December 2013.232.