A Comparison of Feature Selection Techniques for First-day Mortality Prediction in the ICU*

Jacob R. Epifano, Alison Silvestri, Aakash Tripathi, Alexander Yu, Ghulam Rasool, and Ravi P. Ramachandran

Rowan University, Glassboro NJ 08028, USA
{epifanoj0, silves55, tripat67, yualex62}@students.rowan.edu
{rasool, ravi}@rowan.edu

Abstract. The application of machine learning techniques in healthcare has been a growing area of interest around the world. In this paper, we expand on our previous work of mortality prediction by considering various feature sets chosen by several feature selection methods. These methods were implemented and then analyzed. The area under the Receiver Operating Characteristic curve (ROC AUC) was used to evaluate each feature set. The results of each feature selection method were compared to each other as well as our previous analysis. We found that the set of optimal features differed significantly from clinician opinion when a wider feature set was available. We found that Elastic Net was the overall best performing method and was able to reach the same performance as our previous analysis with less than half the features.

1 Introduction

The use of machine learning in today's healthcare systems is rapidly growing as many new techniques are being developed with applications including predicting illnesses, determining the most effective treatments, making quicker and more accurate diagnosis, and many more [7]. The growing demand for more personalized healthcare is becoming an increasingly pressing need. It is imperative that the methods created today are precise in their input. Until Electronic Health Record (EHR) integration can be achieved, we can not expect users to enter large swaths of data for each prediction. The following research will focus on utilizing machine learning and various feature selection techniques in order to more efficiently and accurately predict mortality in the Intensive Care Unit (ICU).

Feature selection is the process of reducing the number of input variables such that the most relevant attributes are used to create a predictive model [12]. There are a total of 194 features within the eICU database [11], therefore it is imperative that this number is greatly reduced to not burden the user. Having excess tests, or features, can not only slow down the time it takes to collect

* Jacob R. Epifano is supported by US Department of Education GAANN award P200A180055. Ghulam Rasool was partly supported by NSF OAC-2008690.

and enter the data, but it can also have a negative impact on the results of the prediction [4]. Having the ability to decrease the number of features used while maintaining a relative measurement of success is crucial for this experimentation and is the motivation behind using feature selection for this project.

Feature selection can be further classified as either supervised or unsupervised. In an unsupervised method, the function is provided inputs, but no target function. The goal is to identify the most relevant features in the input data set. This can be achieved by calculating the correlation with the target variable, or by applying feature-specific statistical tests (such as t-tests). Supervised methods, on the other hand, have the goal of identifying the most relevant features in order to improve the performance of a target function. This can be done by using a training dataset, where the target output is known. The features that correlate with the target variable are identified, and used to optimize the target function. In this work, the filter, wrapper, and embedded methods under supervised feature selection will be utilized. Figure 1 shows a hierarchy of feature selection.

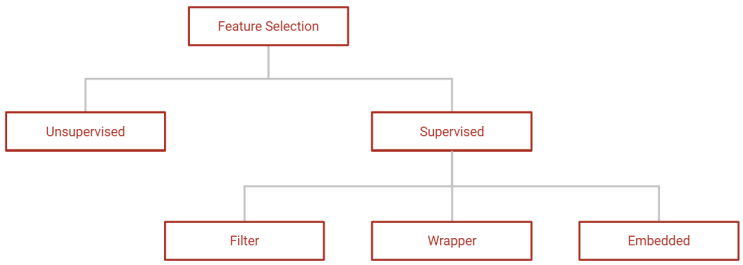

Fig. 1. Feature Selection Hierarchy [2, 8]

The first feature selection method that will be discussed is the filter method. Filter methods are those that are typically performed during the preprocessing stage of feature selection. For this method, the features are chosen based on certain characteristics or scores that they achieve in various statistical tests. Some of the most popular examples of filter methods include correlation, Chi-Square tests, and analysis of variance, also known as ANOVA [9]. The features are then sorted based on their scores and a threshold is chosen either by heuristic or statistics to choose the final set of features.

Wrapper methods are those that evaluate various subsets and combinations of features in order to choose the one that produces the best result for the given algorithm. These methods are often referred to as greedy due to the fact that they search through all subsets and therefore can become computationally expensive and take a long time to execute. The benefit of this method is that it

is guaranteed to provide the optimal set of features if it can run to completion. The process of utilizing a wrapper method starts with choosing a search method or technique in order to select an available subset of features. Once the subset is identified, the desired machine learning algorithm is trained on the chosen subset. The model is then evaluated and the process is repeated with various other subsets of features until the best model is identified. Popular search methods include forward selection, backward elimination, and bi-directional searches [2].

The last feature selection method discussed for this project is the embedded method. As the name suggests, the feature selection is embedded within the training of the model. In the previous two methods, feature selection was performed before the model trained. The process of the embedded method includes training a machine learning model and then deriving the feature importance from the model. As the model is being trained, the features that have the least impact on the prediction are discarded. Examples of embedded methods include lasso/ridge regression and decision trees [2].

The dataset used is retrieved from the eICU database [11]. eICU is a synthesis of data collected from many critical care units throughout the United States. We primarily pull from the *lab.csv* sub-file of this database. This large database included 194 different features of different types, such as: lab tests, demographic data, disease/disorder indicator variables, and vital signs. For this analysis we collected 145,000 instances along with their mortality outcome from the *apachePatientResult.csv* sub-file.

2 Prior Work and Problem Statement

Our work is an expansion on previous work [5]. In that work, a set of 20 features were chosen by an ICU clinician. This work seeks to expand on this by searching for a new set of features that increases the performance of the model while decreasing the number of chosen features. The performance metrics retrieved from [5] will be used as a standard for comparison. The metric we will focus on is the area under the Receiver Operating Characteristics (ROC AUC) and 95% confidence interval obtained from 10-fold cross validation. The AUC is a quantitative measure of how well a model can rank instances of two classes. An AUC of 1.00 would indicate that all negative instances are predicted with a lower probability than positive instances. An AUC of 0.5 indicates that the prediction probabilities are randomly ordered for each class. In this case, the two classes are the mortality results, whether the patient survived or did not survive.

3 Methods

Our analysis consists of three components. First, an evaluation of the dataset with all 194 features. Second, a 1-1 comparison with our prior work by selecting the top 20 features. Lastly, a quantitative and qualitative study of the performance of each feature selection method. The following techniques were evaluated:

variance threshold, analysis of variance (ANOVA), mutual information (MI), recursive feature elimination (RFE), elastic net, and principle component analysis (PCA).

3.1 Variance Threshold

The variance threshold method is a simple, baseline approach to feature selection. It removes any feature that has a variance less than a chosen threshold. By default, this technique eliminates features with zero variance, but can be changed to any desired threshold value. Features with little to no variance may often not contain much significant information and can in some instances reduce the overall performance of a model. Since this method is dependent on the statistics of the feature, thresholding must be done before standardizing the data [10].

3.2 Analysis of Variance

The Analysis of Variance method, often referred to as ANOVA, is a widely popular filter method used during feature selection. For this project, the ANOVA f-test statistic was utilized. However, it is important to note that another form is the ANOVA mutual information statistic. The ANOVA technique is used to determine how similar or different the means of two or more variables are to each other. It can also be utilized to realize the correlation between an independent variable and the dependent variable. Using the ANOVA method, f-scores are assigned to each feature where the higher the f-score indicates a higher correlation between that particular feature and its impact on the output. Therefore, this method can be used to determine which features have little to no impact on the mortality output, and can be discarded from the feature set without having significant effects on the ROC AUC value [9].

3.3 Mutual Information

Shannon Mutual Information between two random variables is defined by conditional entropy. Entropy of the class variable, Y, is desired to be very low in order to maximize classification performance. For a given feature X, Mutual Information between X and Y is a measure of the change in entropy of Y due to the presence of X. Therefore, a high Mutual Information between a feature and the class label indicates that the feature is a good predictor of the class label. [1]

$$I(X;Y) = \sum_{x \epsilon X} \sum_{y \epsilon Y} p(xy) \log \left(\frac{p(xy)}{p(x)p(y)} \right) \tag{1}$$

3.4 Recursive Feature Elimination

Recursive Feature Elimination (RFE) is a wrapper-type feature selection algorithm. It operates by first building a model on a data-set containing all features

Table 1. Statistics for classifier trained on all 194 features.

Measurement	**Mean**	**Standard Error**
Accuracy	0.855106	0.002542
Precision	0.260902	0.003493
Sensitivity	0.819577	0.008339
Specificity	0.857289	0.002884
ROC AUC	**0.918066**	**0.002748**
PRC AUC	0.507141	0.009412
Balanced Accuracy	0.838433	0.003765

and then computing an importance score for the features using the model (ROC AUC for our case). Next, a set of features are removed, and the model is re-trained and outputs an updated importance score on the reduced feature set. This process is then repeated recursively until all the least important features, determined by the importance score from the model, are eliminated. RFE requires two inputs to operate; number of k features to keep and an estimator model with a built-in importance score [10].

To find the k number of features to keep, RFE with cross validation (RFECV) can be utilized. RFECV keeps track of a score computed from the model for a given number of features. Using logistic regression as the model for RFECV and tracking the ROC AUC score, the number of features to keep that meet the Previous Work was found [6].

The following parameters were used; a five-fold stratified for cross-validation, logistic regression as the estimator and the area under the receiver operating characteristic curve (ROC AUC) for scoring the performance of the estimator. The data-set was divided into 33% testing (for scoring) and the remaining was used for training the estimator.

3.5 Elastic Net

Elastic net is an embedded type feature selection algorithm that combines LASSO and RIDGE regression. Both LASSO and RIDGE Regression use regularization to avoid overfitting. Regularization achieves this by penalizing complex models by adding a penalty to the following cost function in Equation 2.

$$W = \sum_{i=1}^{N} \left(y_i \; - \sum_{j=0}^{M} w_j \; x_{\,\mathrm{ij}} \right)^2 \tag{2}$$

RIDGE regression uses L2 regularization, which modifies the cost function by adding a penalty term to the residual sum of squares. Equation 3 shows the added penalty term of lambda times the sum of square of weights.

$$W = \sum_{i=1}^{N} \left(y_i \; - \sum_{j=0}^{M} w_j \; x_{\,\mathrm{ij}} \right)^2 + \lambda \sum_{j=0}^{M} w_j^2 \tag{3}$$

Table 2. Comparison of feature selection techniques. The top 20 features are chosen from each method for a direct comparison of the clinician chosen features from the previous work. The metric reported is the mean ROC AUC from the 10-fold cross validation as well as the 95% confidence interval. Bolded row is the best performing method at 20 features.

Method	ROC AUC
Clinician Chosen [5]	0.8564 ± 0.002748
Variance Threshold	0.7661 ± 0.00747
ANOVA	0.8901 ± 0.00309
Mutual Information	0.8365 ± 0.00554
PCA	0.8980 ± 0.00302
ElasticNet	$\mathbf{0.9056 \pm 0.00336}$
Recursive Feature Elimination	0.8512 ± 0.00415

For Least Absolute Shrinkage and Selection Operator or LASSO regression L1 regularization is used, which modifies the cost function by adding a penalty term of lambda times the sum of absolute value of weights to the residual sum of squares. This is shown in Equation 4.

$$W = \sum_{i=1}^{N} \left(y_i \; - \sum_{j=0}^{M} w_j \; x_{\,\mathrm{ij}} \right)^2 + \lambda \sum_{j=0}^{M} |w_j| \tag{4}$$

Finally, when combining the two penalty terms from LASSO and Ridge in Elastic-Net, an additional α term is added that determines the ratio of L1 to L2 regularization. The Elastic-Net cost function is given in Equation 5.

$$W = \sum_{i=1}^{N} \left(y_i \; - \sum_{j=0}^{M} w_j \; x_{\,\mathrm{ij}} \right)^2 + \alpha\lambda \sum_{j=0}^{M} |w_j| + (1-\alpha)\,\lambda \sum_{j=0}^{M} w_j^2 \tag{5}$$

The values of α and λ for the elastic net equation can be computed using elastic net with cross validation (ElasticNetCV) function. The following list of α was provided and the value of λ was picked automatically from the ElasticNetCV function [10]. α was chosen by performing grid search on the interval $[0, 1]$. The following parameter values were found to be optimal: $\lambda = 4.317\text{e-}4$, $\alpha = 0.995$.

3.6 Principal Component Analysis

Principal Component Analysis, more commonly known as PCA, is an unsupervised method that can be used for dimensionality reduction while having a maximum variability. The overall goal is to make sure all of the of the transformed features are linearly independent, as well as, finding components in order of highest importance. PCA can be defined as the eigendecomposition of the covariance matrix $\mathrm{X^T X}$.

Table 3. Top 10 features selected by each method as compared to Clinician opinion. FR - Feature Rank.

FR	Clinician Survey [5]	Variance Threshold	ANOVA	Mutual Information	Recursive Feature Elimination	Elastic Net
1	Age	CPK	Lactate	Lactate	Potassium	Lactate
2	Immunosuppresion	AST (SGOT)	GCS Motor	Mechanical Ventilation	paCO2	GCS Motor
3	HepaticFailure	Lymphs	GCS Eyes	GCS Eyes	pH	paCO2
4	Lactate	ALT (SGPT)	GCS Verbal	GCS Motor	Myglobin	Bicarbonate
5	Metastatic Cancer	FiO2	Intubation	GCS Verbal	Lactate	Fio2
6	Creatinine	Glucose	Ventilator	Albumin	Glucose	BUN
7	Leukemia	Platelets	AST (SGOT)	Age	Respiratory Rate	WBC
8	Platelet	Akaline	PT-INR	Creatinine	Chloride	Mechanical Ventilation
9	Bilirubin	PaO2	Anion Gap	BUN	Albumin	Sodium
10	Aids	Glucose	ALT (SGPT)	INR	TV	Age

There are five steps to complete the process of principal component analysis. To begin with, the data must be standardized using Equation 6.

$$x_{\text{stand}} = \frac{x - \mu}{\sigma} \tag{6}$$

where x is the original data, μ is the mean, and σ is the standard deviation. The standardized data would then have a mean of 0 for each feature while having a standard deviation of 1. Doing this will scale all features properly and prevent skewing in the results. Step 2 involves finding the covariance matrix of the standardized data. Equation 7 was used to find the covariance matrix where $\bar{x}_i$ is the mean of the ith column, $\bar{x}_j$ is the mean of the jth column, x_{im} is the ith column and x_{jm} is the jth column.

$$\text{Cov}\,(i,j) = \frac{1}{n-1} \sum_{m=1}^{n} \left(x_{\text{im}} - \bar{x_i}\right) \left(x_{\text{jm}} - \bar{x}_j\right) \tag{7}$$

The resultant covariance matrix will be a square matrix $\mathrm{X^T X}$. The next step was to find the eigenvectors and eigenvalues using eigendecomposition. After finding the eigenvalues, the fourth step is to sort them and its corresponding eigenvectors in descending order. The fifth and final step is to choose the amount of components to keep from the highest importance to then transform the standardized data into a transformed matrix using Equation 8.

$$T_R = XW_R \tag{8}$$

where T_R is the transformed matrix, W_R are the loadings or eigenvectors, X is the standardized data, and R itself is the number of components chosen.

4 Experimental Protocol

For our first and second analyses, features with missingness at 70% or greater were dropped from the dataset. Next, features with Pearson correlations greater than 0.9 were also discarded. The data was then standardized and imputed with multiple imputation. To correct for the 95%-5% class imbalance, we perform Synthetic Minority Oversampling (SMOTE) technique [3]. For each 20-feature

Table 4. Minimum number of features required to have statistically significant performance increase over previous work

Method	Number of features
Variance Threshold	102
ANOVA	13
Mutual Information	30
PCA	3
ElasticNet	6
Recursive Feature Elimination	26

set, we perform 10-fold cross validation and collect the mean ROC AUC and 95% confidence interval. Our classifier is inherited from our prior work. We use a multi-layer perceptron (MLP) with 2 hidden layers and SELU activations. The model is trained for 127 epochs using SGD with Nesterov momentum. The hyperparameter values were obtained previously by performing Bayesian Optimization using ROC AUC as the target metric [5].

For our third analysis, we want to consider all features, therefore we do not drop features based on missingness and correlation. For each method, we perform the selection for an increasing number of features and perform 10-fold cross validation to collect a ROC AUC and 95% confidence interval. The rest of the preprocessing procedure remains the same. After training, for each method, we have 194 ROC AUC means and 95% confidence intervals. To quantitatively measure which method performs the best, we can compute the area under the ROC AUC vs number of features curve. In addition, for each method, we can find the minimum number of features that can provide a statistically significant result over the prior work. Lastly, we provide a qualitative measure of performance by comparing the top selected features from each method compared to clinician opinion [5]. Our baseline measure of performance is to consider using all features. The performance metrics for this baseline can be found in Table 1. A link to our code has been provided [1].

5 Results and Discussion

In Table 1 we show the target performance by using all 194 features. Next we cut down the number of features to what was used in our previous analysis which is 20 features. The results for each method is shown in Table 2. Here we can see the top methods are ANOVA, PCA and ElasticNet. When looking at the number of features required to beat the clinicians, in Table 4, we show that PCA and Elastic Net vastly out perform other methods. This trend continues in Figure 2. PCA is quick to rise to a ROC AUC of 0.89 but has trouble increasing as more components are added. The best performing method overall is Elastic Net due to it having the highest AUC as features are increased as well as consistent performance across the other experiments.

[1] `https://github.com/jrepifano/FeatureSelectionInTheICU`

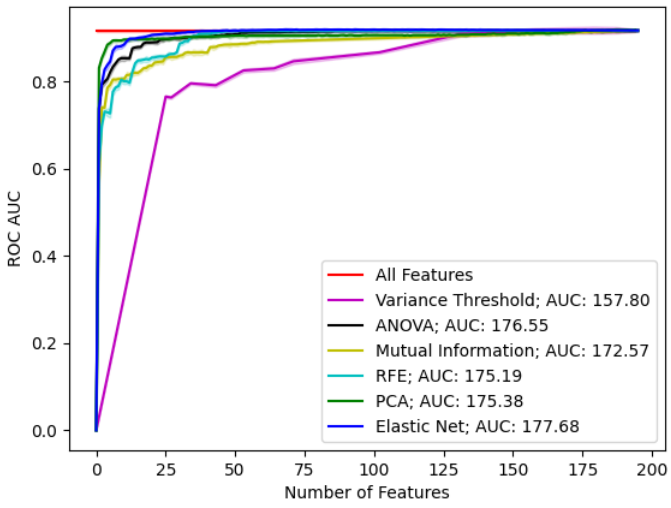

Fig. 2. 10 fold cross validated ROC AUC as a function of number of features selected. The best performing method will have the highest area under the curve.

In Table 3 we show the top 10 features chosen by each method and compare them to what is selected by ICU clinicians. In the clinician column, we observe a large amount of indicator variables as opposed to the columns selected by feature selection. There are also many common features/feature sets that are selected by multiple methods. These include: lactate, GCS scores (Eyes/Motor/Verbal), mechanical ventilation, and age. We postulate that this difference can be attributed to clinician focus on comorbidities with our machine learning models having no knowledge of diagnosis and patient history. It is possible that the optimal feature set lies somewhere in the combination of these methods. We defer the use of a meta set of features selected by multiple methods for future work.

6 Summary and Conclusions

In our study, we provide a quantitative and qualitative assessments of feature selection methods on the eICU first day mortality dataset. We conclude that Elastic Net is the overall top performing method and requires the fewest number of features to match clinician performance while retaining the original features. Finally, we conclude that there is a significant difference in the types of features chosen by humans vs machines. Humans tend to choose features that indicate the presence of disease or comorbidities and the machines tend to choose continuous lab based features. We postulate that the optimal feature set may lie in the intersection of the sets of features selected by these various methods.

Bibliography

[1] Brown, G.: A new perspective for information theoretic feature selection. In: van Dyk, D., Welling, M. (eds.) Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research, vol. 5, pp. 49–56. PMLR, Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA (16–18 Apr 2009), http://proceedings.mlr.press/v5/brown09a.html

[2] Chandrashekar, G., Sahin, F.: A survey on feature selection methods. Computers & electrical engineering **40**(1), 16–28 (2014)

[3] Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: Synthetic minority over-sampling technique. The Journal of artificial intelligence research **16**, 321–357 (2002)

[4] Chu, C., Hsu, A.L., Chou, K.H., Bandettini, P., Lin, C.: Does feature selection improve classification accuracy? impact of sample size and feature selection on classification using anatomical magnetic resonance images. NeuroImage (Orlando, Fla.) **60**(1), 59–70 (2012)

[5] Epifano, J.R., Ramachandran, R.P., Patel, S., Rasool, G.: Towards an explainable mortality prediction model. In: 2020 IEEE 30th International Workshop on Machine Learning for Signal Processing (MLSP). pp. 1–6. IEEE (2020)

[6] Guyon, I., Weston, J., Barnhill, S., Vapnik, V.: Gene selection for cancer classification using support vector machines. Machine learning **46**(1), 389–422 (2002)

[7] Hunter, P.: The advent of ai and deep learning in diagnostics and imaging: Machine learning systems have potential to improve diagnostics in healthcare and imaging systems in research. EMBO reports **20**(7), e48559– (2019)

[8] Miao, J., Niu, L.: A survey on feature selection. Procedia Computer Science **91**, 919–926 (2016)

[9] Miller Jr, R.G.: Beyond ANOVA: basics of applied statistics. CRC press (1997)

[10] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)

[11] Pollard, T.J., Johnson, A.E.W., Raffa, J.D., Celi, L.A., Mark, R.G., Badawi, O.: The eicu collaborative research database, a freely available multi-center database for critical care research. Scientific Data **5**(1) (2018). https://doi.org/10.1038/sdata.2018.178

[12] Shilaskar, S., Ghatol, A.: Feature selection for medical diagnosis : Evaluation for cardiovascular diseases. Expert systems with applications **40**(10), 4146–4153 (2013)