Finding an Optimal Model for Prediction of Shock Outcomes through Machine Learning

Sharad Shandilya
Xuguang Qi
Kayvan Najarian
Rosalyn H Hargraves
School of Engineering
Virginia Commonwealth University
Richmond VA
shandilya.sharad@gmail.com

Michael C Kurz
Department of Emergency Medicine,
Virginia Commonwealth University,
Richmond, VA
mkurz@vcu.edu

Kevin R Ward
Department of Emergency Medicine,
Michigan Critical Injury and Illness Research Center,
University of Michigan,
Ann Arbor, MI
keward@med.umich.edu

Abstract—**Predicting defibrillation success is of paramount importance to resuscitating a victim of cardiac arrest. Several studies have attempted to extract/discover predictive features from electrocardiogram signals. Till date, no method has been accepted or implemented in the field, primarily due to low accuracy and/or specificity. We process a relatively large database of signals and report performance of an integrative Machine Learning model through multiple measures. 358 signals, with 140 leading to return of spontaneous circulation through defibrillation attempts and the rest, 218 signals, leading to unsuccessful defibrillations were used to train and test the model on non-overlapping sample sets. Techniques from machine learning, non-linear dynamics and signal processing were applied to extract features and subsequently classify them. In this study, we identify opportunities for reducing variance in the predictive model and propose a method for searching the optimal model. The accuracy and Receiver Operating Characteristic area of the proposed model are 78.8% and 83.2%, respectively. These compare with 74% and 69.2% accuracy and Receiver Operating Characteristic area for the leading 'Amplitude Spectrum Area' measure. The performance of the model will further improve with addition of other physiologic signals, as previously shown in a study by our research group. The model shows great potential to be viable in the clinical setting.**

Keywords-predictive model; overfitting; machine learning; defibrillation success; parameter search.

I. INTRODUCTION

In the United States, 300,000 lives are annually lost due to cardiac arrest. Survival rates for out-of-hospital cardiac arrest patients are very low [2]. Ventricular Fibrillation (VF) is a common arrhythmia in cardiac arrest [3]. Coronary artery perfusion provided by Cardio-Pulmonary Resuscitation (CPR) prior to defibrillation has been shown to improve chances for Return of Spontaneous Circulation (ROSC) [4]. Defibrillation is a procedure that delivers an electrical current that depolarizes a critical mass of the myocardium simultaneously. Defibrillation increases the possibility of the sino-atrial node regaining control of the rhythm. Coronary artery perfusion provided by CPR prior to defibrillation has been shown to improve chances for ROSC [4]. Repetitive unsuccessful shocks can reduce chest compression time and cause injury to cardiac tissue, impacting heart function upon survival. Even worse, unsuccessful shocks can cause VF to deteriorate into asystole or Pulseless Electrical Activity (PEA), which are more difficult to resuscitate [5]. A victim's chances of survival worsen by 10% for every minute of VF that remains untreated [4].

Hence, increasing efficacy of defibrillation attempts is of principal importance. To achieve this, we develop an integrative decision-support model that guides the interventionist by learning from real-time information extracted from the patient.

Fourier Transform (FT) based methods [1] assume a linear and deterministic basis for decomposing signals. As another limitation, FT decomposition yields a time averaged frequency estimate of the original signal. These limitations, except for the *deterministic* and *non-chaotic* assumption, are overcome by our use of the Wavelet Transform (WT) [6]. Furthermore, we use a dual-tree decomposition algorithm for the complex WT, nearly eliminating shift-variance, which is a limitation of DWT.

Additionally, the QPD-PD method [6] is able to characterize chaotic signals while allowing for stochasticity/non-determinism. In the same study, it was shown that features calculated through QPD-PD, along with those from WT, represent a powerful set whose knowledge can be integrated with a Machine Learning (ML) algorithm for improved performance.

Furthermore, we propose a novel method of selecting the optimal ML model to boost generalize-ability on blind data.

II. OBJECTIVES

Thus far, no automated model or feature has been accepted in the field for decision-support during cardiac arrest due to low specificity at desired sensitivity and/or low overall accuracy. By relaxing the assumptions made by methods previously tried, as described in the previous section, we aimed to build a model with high sensitivity and specificity. Model (parameter) selection is another important endeavor in boosting performance, especially when employing heuristic-based ML algorithms.

 ISBN: 978-1-61208-283-7

A high specificity translates to a reduced number of unnecessary shocks, which cause thermal injury to the heart in addition to adding to the time lost.

Section 3.1 gives an overview of the methods and the novel techniques proposed. Section 3.2 details the data used to build and test the model. Section 3.3 pertains to filtering as a data pre-processing step. Section 3.4 describes the QPD-PD method used to characterize data prior to calculation of features. Section 3.5 describes the proposed method for model selection. Sections 3.6 reports how the model's performance was tested. Section 4 contains results of a comparative study with the leading Amplitude Spectrum Area (AMSA) measure [1]. Section 5 states the conclusions from this study.

III. METHODOLOGY

3.1 Overview

Data was characterized through techniques from non-linear dynamics, autorecursive modeling, time-frequency decomposition. Novel time-series features were devised in order to distinguish pre-defibrillation VF signals yielding a successful defibrillation from those that did not. The method Quasi-Period Density Prototype Distance (QPD-PD) derives stochastic quasi-periods through time-delay embedding. Supervised feature selection was performed to identify the most discriminative features. Selection was performed in a nested fashion so as to maintain blindness to the test folds. In the same nested fashion, model selection was performed for different combinations of parameters. The overall approach is discussed in [6]. In this paper, we propose a novel method of model selection and term it "High-Platform Method". Additionally, the assumptions that underlie the QPD-PD model are formally tested and reported. Model performance is reported on a newly acquired large (for the problem context and when compared to other studies) dataset provided by Zoll Medical Corporation [Chelmsford, MA]. Simultaneous 10-fold cross-validation was used to evaluate the model. Matlab® software was utilized for all signal-processing needs. Figure 1 illustrates the high-level steps of the methodology.

Blocks A1, A2 and B1, B2 represent pre-processing of signals. C1, C2, and C3 represent extraction of features/characteristics from the data/signals. Blocks D1 and D2 represent the ML phase of the system where features are selected, model selection is performed by the proposed method, and the inducted ML model is tested.

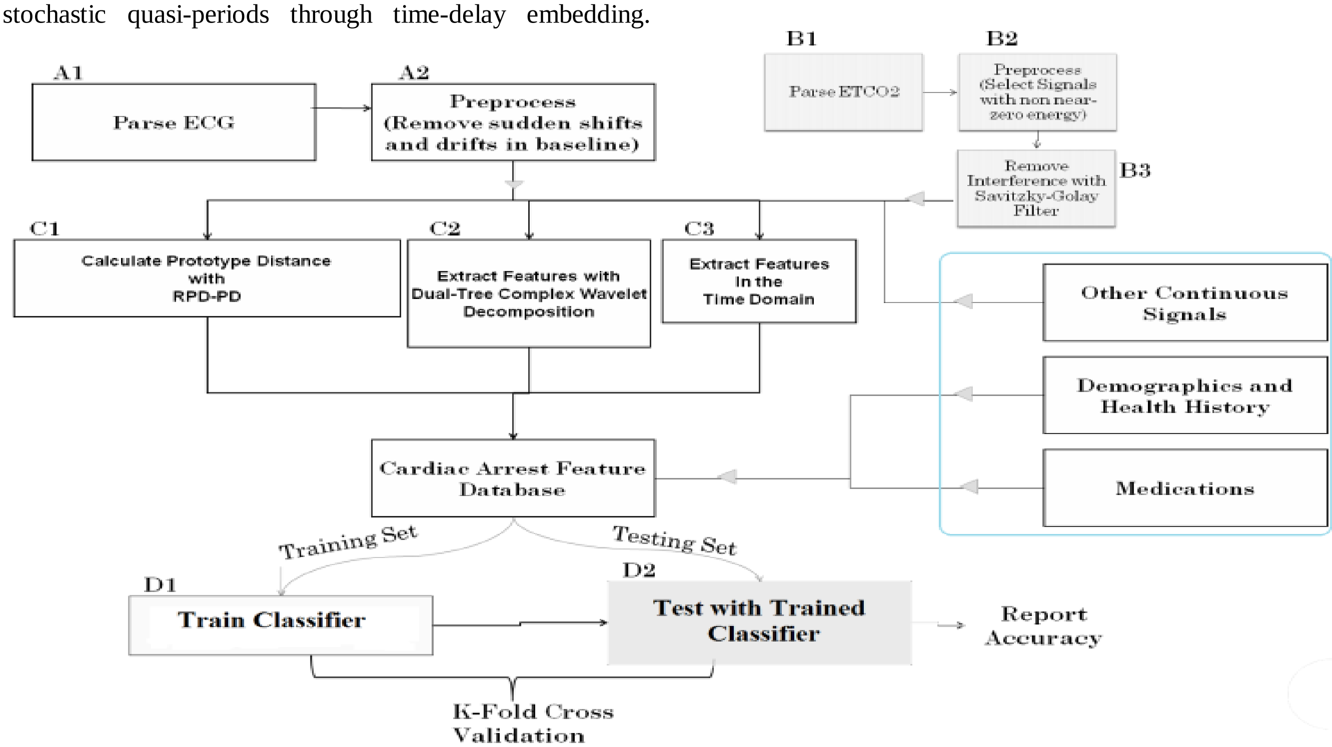

Figure 1. Overview of the System

3.2 The Data

The data processed in this study contained a total of 358 defibrillation counter-shocks on 153 subjects. Every single one of these shocks was labeled / annotated by Dr. Michael Kurz. Among these shocks, 140 were successful and 218 were unsuccessful. The same dataset was used for training and testing of both VCU algorithm and the AMSA method as described by [1]. Successful defibrillation was defined as a period of greater than 15 seconds with narrow QRS complexes under 150 beats per minute with confirmatory evidence from the medical record or electrocardiogram

(ECG) that a return of spontaneous circulation (ROSC) has occurred.

3.3 Pre-Processing

Signals were filtered by utilizing the method proposed in [7]. The method performs custom filtering and was designed by observing properties of the ECG signals in the database.

3.4 Feature Extraction

This QPD-PD method focuses on distributions of pseudo-periodicity, while accounting for stochastic character of the signal. Parameter selection and feature calculation are geared for classification. Figure 2 shows the quasi-period density plots for each class. The periods were then convolved with the exponential function in order to quantify the difference between the two densities. For the density represented by *Q*, we propose to calculate a new probability density function for Q by convolving as follows

$$\mathrm{Q} * \mathrm{Exp}\,(\mathrm{s}) = \int_0^{\mathrm{P}} \mathrm{Q}(\mathrm{p})\,\mathrm{Exp}(\mathrm{s}-\mathrm{p})\,\mathrm{dp} \qquad (1)$$

where *Exp* is the exponential function, *s* is the shift, *p* is a specific quasi-period, *P* is the largest period recorded. Dual-tree complex wavelet transform and other time-series features were also calculated as described in [6].

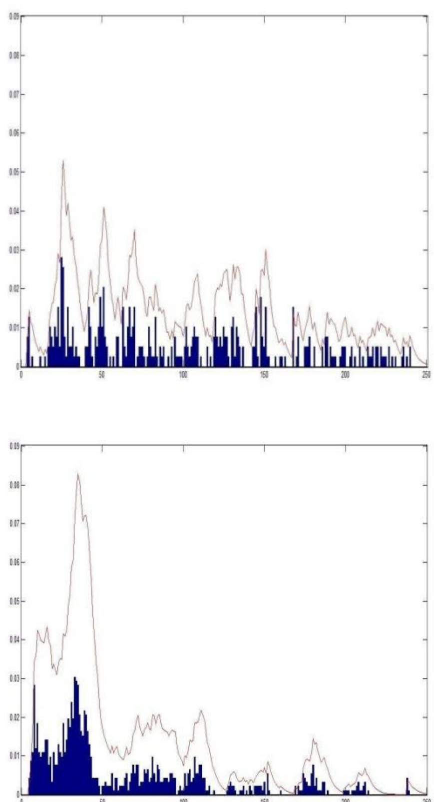

Figure 2. Quasi-Period Density. X-axis: Quasi-Periods; Y-axis: Amplitude Density. QPD for a successful shock (above) and QPD for an unsuccessful shock (below). Blue bars represent the normalized amplitude for each pseudo period. Red line represents QPD convolved with the exponential function. If most of the amplitude is clustered in neighboring Quasi-Periods, as is the case above, the convolution helps accentuate that fact (higher peak in the line-plot).

3.5 Feature and Model Selection

Feature selection, performed with cross-validation on the whole dataset, creates a positive bias in accuracies by indirectly using information from the test set. As such, feature selection must be performed within the training set that is generated for each run of *k*-fold cross-validation. However, using the entire training set leads to over-fitting within the training set, which creates a negative bias in accuracies when each test fold is passed through the model [8]. To prevent this, and to also select parameters for the learning algorithm in a nested fashion, we employ a twice-nested version of cross-validation.

For parameter tuning, we propose the 'High-platform' method. Figure 3 represents a plot of median accuracy for different combinations of parameters. We assume that close values of parameters create models that are conceptually and performance-wise similar. The conjecture is that picking a model that does both, maximizes the median accuracy and belongs to a 'high-performing neighborhood', would preclude an overfitted model that may have a high accuracy, but would be adjacent to other poorly performing models. After combinations of all parameters are tried and median-accuracies are recorded, we pick a model that

1) Exists within the neighborhood that has the highest mean median-accuracy, and
2) Has the highest median-accuracy within that neighborhood.

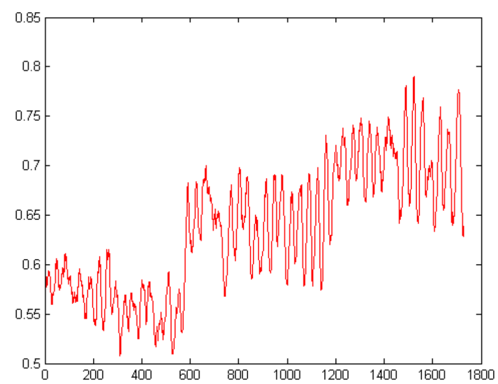

Figure 3. Finding a High-Platform. Four parameters (Learning Rate, Momentum, Hidden Neurons, and Epochs) for a Neural Network are varied. X-axis: Unique combinations of parameters. Y-axis: Median cross-validated accuracy for each combination. A region, such as the one between 1200 and 1450, with the highest mean median-accuracy is chosen.

Each neighborhood is defined by a fixed set of values for the subset of parameters that we want to optimize. Then averaging the accuracy over a neighborhood yields the 'platform', which amounts to nullifying the effect of varying values of the remaining parameters from the superset. For instance, optimizing a total of four parameters would involve the following. After all combinations of possible values of the 4 different parameters are tried,

1) Calculate the average accuracy for each unique combination of the first three parameters
2) Find and fix the combination that has the highest average accuracy,
3) Then, vary values of the fourth parameter and select the model with the highest accuracy.

To aid comprehension, this is similar to a *pseudo* "best-first" approach for parameter tuning while keeping the order of parameters the same as above. In this procedure, an exhaustive search would be performed for two parameters, then the third and fourth are chosen one at a time. It would progress as follows:

1) Try all possible combinations of the first two parameters, fix the best one, and call it Opt2 ,
2) Try all values for the third parameter (for a fixed value of the fourth parameter) and note the best one,
3) Try all values of the fourth parameter (for a fixed value of the third parameter) and note the best one,
4) From steps 2 and 3 above, pick the model with higher accuracy and call it Opt3,
5) Optimize on the remaining parameter, yielding Opt4.

This is similar to, but not exactly, best-first search because values for the third and fourth parameters are selected apriori during the procedure. Instead, the proposed high-platform method searches for the best model by taking one or more steps back from exhaustive search. For one-step-back, it reduces variance at both, penultimate and ultimate levels. Optimizing at the penultimate level is done by averaging variation in performance induced by varying values of the last parameter. For a greater reduction in variance, the search would take two-steps-back and average the variation in performance induced by all combinations of the remaining *two* parameters.

3.6 Classification and Validation

Cross-validation (CV) is a ubiquitous method that creates non-overlapping training and testing sets from the data set in order to train and test a given model. It is essential in avoiding a positive bias in accuracy when the entire data set is used for both training and testing. CV is desirable because it allows utilization of the entire dataset in order to validate a method, as opposed to having a single training set and a single test set. Multiple comparisons of the proposed model and AMSA method [1] were performed using 10-fold cross validation.

IV. RESULTS AND DISCUSSION

Overall accuracy for the proposed model was 78.8%. We compared the proposed model with the current state-of-the-art AMSA method as follows.

- Comparison at 80% sensitivity: In this study, the two algorithms, proposed model and AMSA, were trained to provide sensitivity of 80%. In this case, our model provided an accuracy of 74% and specificity of 70.2%. For the same level of sensitivity, AMSA provided an accuracy of 53.6% and specificity of 36.7%.
- Comparison at 90% sensitivity: A similar analysis was conducted, except that both algorithms were trained to provide a sensitivity of 90%. our method provided an accuracy of 68.4% and specificity of 54.6%. For the same level of sensitivity, AMSA provided an accuracy of 43.3% and specificity of 13.3%.
- Using Receiver Operating Characteristic testing, the Area Under the Curve (AUC) for proposed method is 83.2% while this number for AMSA is 69.2%. These results are similar to the ones reported in [6].

For a given desired sensitivity, the proposed model can provide a significantly higher accuracy and specificity. Notably, within the range of 80-90% of sensitivity, the model provides about 40% higher specificity. This means that when trained to have the same level of sensitivity, the model will have far fewer false positives (unnecessary shocks). Furthermore, the significantly higher AUC of the selected model suggests significantly higher reliability than AMSA.

V. CONCLUSION AND FUTURE WORK

We have developed a novel algorithm for predicting successful defibrillation of VF and then selecting an optimal model. The model is built upon knowledge extracted with signal-processing, non-linear dynamical and machine-learning methods. The selected ML model shows viability for decision-assistance in clinical settings. Our approach, which has focused on integration of multiple features through machine learning techniques, suits well to boosting predictive accuracy and model robustness.

Improvements will be sought in feature selection through novel methods. Performance of ML algorithms from disparate paradigms will be compared after model selection with a fixed (selected) feature subset.

REFERENCES

[1] G. Ristagno, A. Gullo, G. Berlot, U. Lucangelo, F. Geheb, and J. Bisera, "Prediction of successful defibrillation in human victims of out-of-hospital cardiac arrest: a retrospective electrocardiographic analysis," Anaesth Intensive Care, vol. 36, 2008, pp. 46-50.

[2] G. Nichol, E. Thomas, and C. W. Callaway, "Regional variation in out-of-hospital cardiac arrest incidence and outcome," J Am Med Assoc, vol. 300, 2008, pp. 1423–1431.

[3] V. M. Nadkarni et al, "First documented rhythm and clinical outcome from in-hospital cardiac arrest among children and adults," JAMA, vol. 295, 2006, pp. 50–57.

[4] T. D. Valenzuela, D. J. Roe, S. Cretin, D. W. Spaite, and M. P. Larsen, "Estimating effectiveness of cardiac arrest interventions: a logistic regression survival model," Circulation, vol. 96, 1997, pp. 3308–3313.

[5] H. Strohmenger, "Predicting Defibrillation Success," Cardiopulmonary Resuscitation, vol. 14, 2008, pp. 311-316.

[6] S. Shandilya, K Ward, M Kurz, K Najarian. "Non-Linear Dynamical Time-Series Characterization for Prediction of

Defibrillation Success through Machine Learning", BMC Informatics and Decision Making, 2012.

[7] S. Shandilya, M. C. Kurz, K. R. Ward, and K. Najarian, "Predicting defibrillation success with a multiple-domain model using machine learning," IEEE Complex Medical Engineering, 2011, pp. 22-25.

[8] R. Kohavi and G. John, "Wrappers for feature subset selection", Artificial Intelligence, vol. 97, 1997, pp. 273-324.

 ISBN: 978-1-61208-283-7