Predicting Sepsis Biomarker Progression under Therapy

Ivan Stojkovic[†‡] and Zoran Obradovic[†]
[†]*Computer & Information Sciences Department, College of Science & Technology*
Temple University, Philadelphia, PA 19122, USA
Email: zoran.obradovic@temple.edu
[‡]*Signals & Systems Department, School of Electrical Engineering*
University of Belgrade, 11120 Belgrade, Serbia
Email: ivan.stojkovic@temple.edu

***Abstract*—Sepsis is a serious, life-threatening condition that presents a growing problem in medicine and health-care. It is characterized by quick progression and high variability in the disease manifestation, so treatment should be personalized and tailored to fit individual characteristics of a particular subject. That requires close monitoring of the patient's state and reliable predictions of how the targeted therapy will affect sepsis progression over time. We have characterized predictive capabilities of a graph-based structured regression approach under hemoadsorption therapy by using a computational model of sepsis biomarker progression in rats. Results suggests that an extension of the model representational power by using a dense graph and multiple-step predictors increases predictive accuracy, allowing more appropriate choice of treatment.**

***Keywords*-biomarker prediction; sepsis treatment; structured regression; time series; computational model;**

I. Introduction

Sepsis is a complication of pathogen infection that triggers an uncontrolled systemic inflammatory response, often resulting in dysfunction of multiple organs and even death [1]. It is a growing problem for modern society, as it affects a large population [2]. Septic shock is a major contributor to in-hospital deaths [3] and incurs significant costs on the healthcare system, with over $20 billion spent in 2011 in the United States alone [4]. Some of the main challenges that impede the successful management of sepsis are heterogeneity of symptoms and fast progression [5]. Necessity for a more effective treatment is apparent, and it seems that precision medicine will be an important part of the solution [6]. Such a personalized approach requires customized therapy based on accurate predictions and careful monitoring of a subject's state in order to track the disease progression and quickly modify the therapy if needed. In this article, we focus on prediction accuracy in the presence of treatment, as it is a prerequisite for deciding an appropriate therapy.

II. Related work

An optimal therapy is the one that would provide maximal benefits in driving the patient towards the healthy state. The optimal effectiveness of two recent blood cleansing approaches for sepsis treatment was recently assessed [7], where a computational model of sepsis was extensively tested under a number of possible therapy configurations, and outcomes were compared to find the best treatment policy. However, this requires repeated application of different therapy configurations under the same initial conditions, while in reality there is only one such event. In practice, the effects of therapy could only be predicted based on previous experience in other cases, and the optimal therapy would be the one with the most favorable expected outcome.

A prediction-based therapy optimization for sepsis, which used a differential equations sepsis model [8] was explored in [9]. Later, the same Data Driven Model Predictive Control (DD-MPC) approach was also applied [10], with a more detailed computational model of sepsis biomarker progression that included cytokines and neutrophils [11].

Even though the proposed DD-MPC framework achieved improved therapy efficiency, it relied on a simple linear regression model for obtaining the prediction of a subject's response to therapy. In order to further improve the efficiency of treatment, more accurate predictions are needed. Therefore, a more sophisticated prediction model for sepsis progression, based on Gaussian Conditional Random Fields (GCRF) [12], was proposed [13], [14].

In this article, we build upon the approach investigated in [13], but using a more complex mathematical model of sepsis dynamics proposed in [11]. We employ the GCRF model for structured regression to predict levels of cytokines, which are important biomarkers for sepsis. Furthermore, we extend the previous efforts by generalizing the approach into two directions. First, in this study we utilize a more general dense graph of possible dependencies, instead of a set of independent linear chains. Second, we explore a direct multi-step-ahead prediction versus a more commonly used approach of iterative one-step-ahead predictions.

We work under the assumption that measurements are certain, always available and without delay, even though the measuring technology is never ideal. In practice, measurements are often sampled at non-uniform intervals [15], and it often takes a non-negligible amount of time from taking samples until the measured values are available. In addition, observations are often missing in practice for some variables, which requires adjustment in methodology [16].

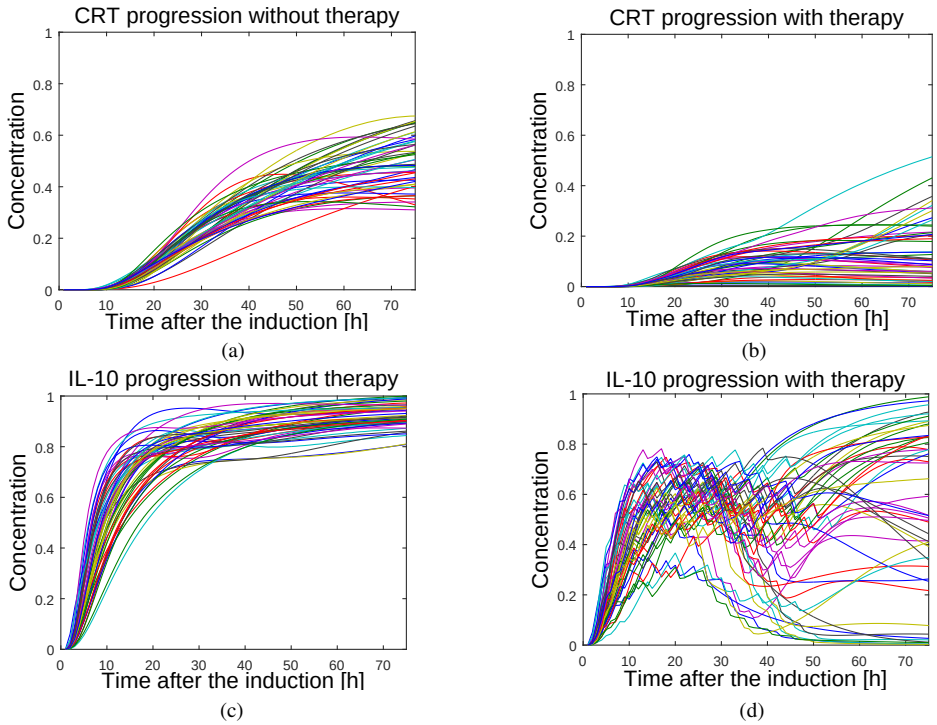

Figure 1: Progression of two cytokines in 30 virtual subjects (a) CRT without therapy, (b) CRT with therapy, (c) IL-10 without therapy and (d) IL-10 with therapy applied.

Also, accounting for the uncertainty in predicted values is shown to be beneficial [17]. Measured data often require proper normalization, to avoid the undesired effects that outliers might introduce [18]. In many applications only the most informative features should be selected [19] in order to focus the computational efforts and to increase prediction accuracy and interpretability of results. Some studies also stress the importance of accounting for disease-disease [20] and disease-gene [21] associations. However, for simplicity, we will not account for nor address these issues in this paper.

III. Approach

A. Model

The computational model that we rely upon in this work is a set of coupled Ordinary Differential Equations (ODE) in the form of a nonlinear system (1) detailed in [7], [11]. Formula f was manually configured using mass action kinetics law, Hill functions and decay factors, based on the known mechanisms of neutrophil trafficking and cytokine dynamics during acute inflammation [11]. The system's state vector $S(t)$ is comprised of 19 time evolving variables, including pathogen levels B resulting from CLP (Cecal Ligation & Puncture) procedure, different types of neutrophils (sequestered N_s, lung N_l, tissue N_t, resting N_r, primed N_p, activated N_a) and cytokines (TNF, Il_{1b}, Il_6, Il_{10}, $Lsel$, $HMGB_1$, CRT, ALT). Additionally there are conceptual variables like total tissue damage D and lumped states of pro- and anti-inflammation (PI and AI). Vector of observations $O(t)$ contains only directly measured states of 8 cytokines (2). An external binary control signal $u(t) \in \{ON, OFF\}$ activates reduction of inflammatory mediators, thus modeling hemoadsorption (HA) treatment [22]. The system of ODEs can be simulated for various values of free parameters θ and under different profiles of external signal $u(t)$ that mimic the effects of therapy.

$$\frac{dS(t)}{dt} = f(S(t), u(t), \theta) \tag{1}$$

$$O(t) = [TNF, Il_{1b}, Il_6, Il_{10}, Lsel, HMGB_1, CRT, ALT]^T \tag{2}$$

B. Data

Free parameters θ of the model (1) were optimized using the Metropolis Hastings algorithm on the real measurements from CLP septic rats [11], so that trajectories closely resemble experimentally observed patterns. We have generated the set of 30 virtual subjects, whose parameters θ^i were sampled from a stationary distribution of the resulting Monte Carlo Markov Chain (MCMC). Trajectories were designed to resemble patterns observed in the death cohort, that is, the rats that would die without a therapy applied. Examples of hourly sampled time-series of cytokines' values for the case where no therapy is applied are given in Figures 1a and 1c.

We have generated training examples by applying 12 doses of hourly HA therapy randomly scattered on the interval of the initial 50 hours since the bacterial infection induction. Examples of how applied therapy influenced observable states are depicted in Figures 1b and 1d.

C. Prediction of biomarker progression

We pose the task of biomarker progression prediction as an estimation of the future biomarker values based on their historical observations. More specifically, measurements (and control signal values) collected up to time t, $X_{t:t-d} = [O_t, u_t, O_{t-1}, u_{t-1}, ..., O_{t-d}, u_{t-d}]^T$ should be used to regress states over the future horizon $Y_{t+1:t+p} = [O_{t+1}, ..., O_{t+p}]^T$, where d is the number of past time-steps and p is the number of future time-steps.

There are two main ways to predict time series over the multiple p time steps in the future horizon. The first way is to predict one step after another by including the previous prediction into the time series, and we will refer to it here as the iterative approach. The other way is to predict each of the future time steps directly from the time series that is observed so far, also known as the direct approach. In limited time series the direct method possesses some theoretical advantages over the iterative one, nevertheless it is not too often advantageous in practice [23]. Because in the case of the iterative method there is only a need to train one step of the prediction model, the iterative approach is preferred over the direct one, where one needs to train multiple models.

Nevertheless, iterative and direct approaches both use the same predictive methods and learning algorithms, and differ mainly in the data that is fed to their inputs. Here we make use of two common machine learning approaches for learning the needed functional mappings.

1) Linear Regression:

$$y_i = W^T X_i + \epsilon, \quad \epsilon\ N(0, \sigma^2) \tag{3}$$

Linear Regression (LR) model, in this particular application, also known as auto regressive model with exogenous input (ARX), assumes that single output y_i depends linearly on its input X_i. It is parametrized through a vector of weights W, which has a closed form solution as a function of training data X and Y.

$$\hat{W} = (X^T X)^{-1} XY \tag{4}$$

Predictions y_{LR} are obtained by weighting the test observations X_t, with $\hat{W}$ learned on the training set.

$$y_{LR} = \hat{W}^T X_t \tag{5}$$

2) Gaussian Process Regression:

$$y_i = gp(m(X), k(X, X')) + \epsilon, \quad \epsilon\ N(0, \sigma^2) \tag{6}$$

Gaussian Processes (GP) are a nonlinear kernel based regression technique that models a finite set of observations as a sample from the distribution over functions. It is specified by a mean function $m(X)$, commonly set to zero without loss of generality, and a positive definite kernel function $k(X, X')$ that essentially determines the properties of the functions [24]. Here we have adopted the superposition of two common kernels (7), linear and exponential, as we observed better generalization properties.

$$k(X, X') = a_0 + a_1 e^{(-a_2 (X - X')^T (X - X'))} + a_3 X^T X' \tag{7}$$

Due to the modeling assumption that any number of samples has a multivariate Normal distribution, inferring the prediction in GP has a convenient analytical formulation.

$$y_{GP} = K^T C^{-1} Y \tag{8}$$

Where $K = k(X, X_t)$ is the kernel matrix of training and testing samples and $C = k(X, X) + \sigma^2 I$ is the kernel matrix of training and training samples augmented with variance on the diagonal elements.

Presented linear (LR) and nonlinear (GP) regression approaches are utilized to learn mappings and give predictions from vector inputs X_i for each of the output variable y_i, and also for different time-steps $p \in \{1, 2, ..., 10\}$, separately. One disadvantage of this approach is that predictions are made somewhat independently of each other, and in fact the values of variables are very dependent between each other and over the time. There exist structured approaches that simultaneously take into account all predictions and correct them to be more consistent. We adopt a Gaussian Conditional Random Fields (GCRF) framework to improve predictions from unstructured Linear Regression and Gaussian Processes based models.

D. Structured Regression using GCRF

Gaussian Conditional Random Fields is a probabilistic graphical framework that models the joint probability function of all variables in target vector Y, given the input X, as an exponentiated weighted sum of association and interaction potentials.

$$Pr(Y|X) = \frac{1}{Z} \exp(-\sum_k \sum_i \alpha_i^k (y_i - R_i^k(X))^2 - \sum_l \sum_{(i,j)} \beta_{ij}^l M_{ij}^l (y_i - y_j)^2) \tag{9}$$

Association potential is a quadratic function of the difference between the estimate of one output y_i and its associated prediction from k-th unstructured model $R_i^k(X)$. Every such potential is weighted with free parameter α_i^k to be learned, which regulates how close the estimate of the variable should be to a particular unstructured prediction. Similarly, interaction potential is a quadratic function that tries to bring closer together the estimates of two related target variables

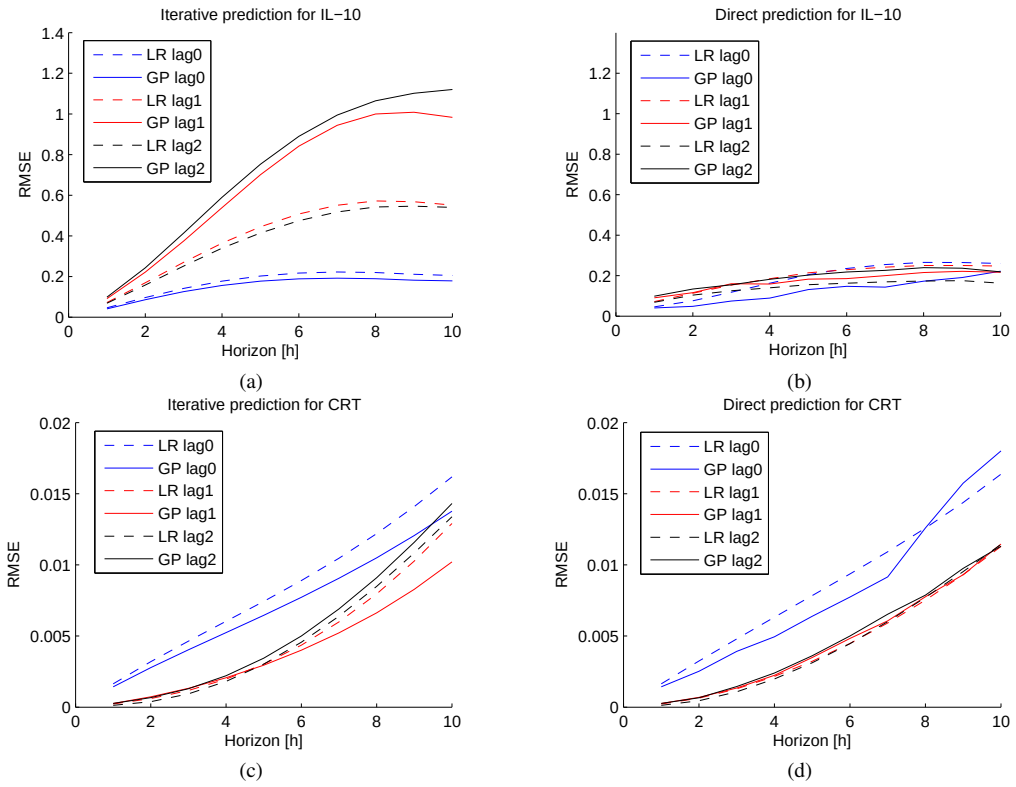

Figure 2: Root mean square error (RMSE) graph for the six unstructured predictors (a) IL-10 iterative approach, (b) IL-10 direct approach, (c) CRT iterative approach, (d) CRT direct approach.

y_i and y_j, where $M^l_{i,j}$ is a connectivity (or similarity) matrix and β^l_{ij} is a free parameter corresponding to the strength of relatedness. Normalization term Z assures that joint probability function $Pr(Y|X)$ is a proper probability, i.e. it integrates to one. Presented version of the GCRF model requires prior learned unstructured predictors $R^k_i(X)$ and similarity matrix M, and free parameters α's and β's can subsequently be learned by minimizing the negative log-likelihood.

The fact that joint probability is expressed as an exponential of a sum of quadratic functions allows expressing the objective in the form of a multivariate Gaussian function. Several convenient properties stem from that insight. Inference has a closed form analytical solution (10), and convexity of the objective allows the use of reliable and efficient optimization algorithms.

$$Y_{GCRF} = Q^{-1}B \tag{10}$$

Where precision matrix $Q = 2(Q^1 + Q^2)$ is composed of two parts Q^1 (the association potential part) and Q^2 (the interaction potential part):

$$Q^1_{i,j} = \begin{cases} \sum_k \alpha^k_i, & \text{if } i = j. \\ 0, & \text{otherwise.} \end{cases} \tag{11}$$

$$Q^2_{i,j} = \begin{cases} \sum_l \sum_{(i,j)} \beta^l_{ik}, & \text{if } i = j. \\ -\sum_l \beta^l_{ij}, & \text{otherwise.} \end{cases} \tag{12}$$

And elements of a vector $B = [b_1, b_2, ..., b_n]^T$ are obtained as a linear combination of the unstructured predictions $R^k_i(X)$:

$$b_i = 2\sum_k \alpha^k_i R^k_i(X) \tag{13}$$

In this way, GCRF prediction Y_{GCRF} simultaneously takes into account individual predictions for different targets y_i and also different predictions R^k_i for a single target, thus behaving as both an ensemble and multitask approach.

IV. Results

The response of some of the observable states to the application of various therapies is shown at Figure 1. Figures 1a and 1c show progression of two cytokines, CRT and IL-10, in virtual subjects which were not on a treatment, while Figures 1b and 1d correspond to the same situations with a therapy applied. It was evident that both states are affected by the therapy. However, predicting in what way HA therapy will affect the progression of IL-10 is a much more difficult problem due to high variability in trajectories (Figure 1d).

In the case of CRT the task was much easier than for IL-10, which can be also concluded from Figure 2, where the performance is reported. In fact, out of the remaining 6 cytokines three of them, TNF, IL-1b and IL-6 follow the pattern of IL-10 and the other three, Lsel, HMGB1 and ALT will be more like CRT. It is worth mentioning that these more difficult states TNF, IL-1b and IL-6 are directly influenced by the unobservable and conceptual Pro-inflammatory state and IL-10 is influenced by the Anti-inflammatory state. On the other hand, easier cytokines for prediction, HMGB1, CRT and ALT are proxies for Tissue Damage, and Lsel is a proxy for activated neutrophils N_a. Since CRT and IL-10 are the two most extreme cases here, we are going to show comparison of the results just for the two of them, as similar results hold for the other cytokines.

First, we compare the two approaches to temporal prediction, indirect and direct, in terms of prediction error. For that purpose we trained the six different unstructured models, with three different lengths of historical observations $d \in \{0, 1, 2\}$, and for two algorithmic approaches LR and GP. Models were trained on 30 virtual subjects' time series data which, along with the 19 states, also included information about the applied therapy. Performances for various setups were accessed on the new test set consisting of 50 virtual subjects. Testing time series included hourly observations up to the 75th hour after the sepsis induction. At each time point, 9 therapy durations were evaluated and predictive models were applied to estimate hourly values of 8 observable states over a 10 hour horizon. The results are shown in Figure 2 on the examples of the two cytokines, CRT and IL-10. In the case of CRT, small improvement in accuracy can be observed in the case of the direct approach, although there is not much difference between the two approaches (bottom panels in Figure 2). For the IL-10 cytokine it is much more emphasized that the direct approach is beneficial for the reduction of the prediction error. It is also interesting that performance of the direct approach on the two cytokines yields comparable accuracy, as can be seen from Figures 2b and 2d.

Second, we evaluate the structured regression GCRF [12] approach that can utilize multiple independent predictions from different predictors (e.g. LR and GP), and for different tasks (e.g. CRT and IL-10), to generate more accurate joint predictions. The GCRF model used in [13], is based on the graph in which the dependency structure consists of 8 separate linear chains with 10 nodes, each node representing one time step in the future horizon. This is a fair assumption since the value of one cytokine should have the strongest connection with the value of the same cytokine one time step before. Now we have relaxed that assumption and have allowed a fully connected graph where the value of each cytokine at each step is connected to all other time steps as well as all other cytokines at all time steps.

The graphics in Figures 3a and 3b are comparing the

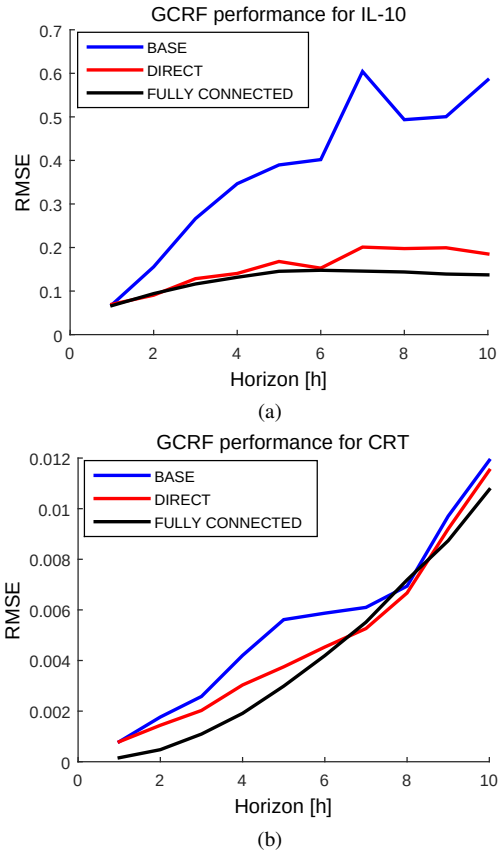

Figure 3: RMSE for GCRF model predictions over 10 hours horizon in case of (a) CRT and (b) IL-10 cytokine, respectively. BASELINE method is GCRF build on top of iterative predictions with linear graph, DIRECT one utilizes direct approach instead, while FULLY CONNECTED corresponds to the case when GCRF contains fully connected graph and is built on top of direct approach.

GCRF performance for the three different cases. Baseline (BASE) is the setup with the linear chain graph and iterative approach in generating unstructured predictors. The obtained errors for this approach are the largest. When we replace the iterative approach with the direct one we get clearly improved results. We refer to that one as DIRECT in the Figures 3a and 3b. When we also replace the linear graph with the "FULLY CONNECTED" one we can observe further reduction in the prediction error.

V. Conclusion

Performed experiments reveal that behavior of some biomarkers in the presence of therapy might be easier to predict, compared to others. Therefore, robust predictive methods are needed in order to perform well for all biomarkers. Based on the results from testing the GCRF model under different structural setups, highly expressive models with more general connectivity structure and with a larger number

of multi-step predictors achieve better performance, in terms of error minimization. Such an increased accuracy could aid in the selection of more appropriate treatments. Potentially, further improvement might be achieved using a structured approach with a more general model of interactions [25].

ACKNOWLEDGMENT

This research was supported in part by DARPA grant FA9550-12-1-0406 negotiated by AFOSR, and in part by DARPA grant No. 66001-11-1-4183 negotiated by SSC Pacific grant. Simulations were performed on the OwlsNest HPC cluster at Temple University, which is supported in part by the National Science Foundation through a major research instrumentation grant (CNS-09-58854).

REFERENCES

[1] R. P. Dellinger, M. M. Levy, A. Rhodes, D. Annane, H. Gerlach, S. M. Opal, *et al.*, "Surviving sepsis campaign: international guidelines for management of severe sepsis and septic shock, 2012," *Intensive care medicine*, vol. 39, no. 2, pp. 165–228, 2013.

[2] Anonymous, "Focus on sepsis," *Nature Medicine*, vol. 18, no. 7, p. 997, 2012.

[3] V. Liu, G. J. Escobar, J. D. Greene, J. Soule, A. Whippy, D. C. Angus, *et al.*, "Hospital deaths in patients with sepsis from 2 independent cohorts," *The Journal of the American Medical Association*, vol. 312, no. 1, pp. 90–92, 2014.

[4] C. M. Torio and R. M. Andrews, "National inpatient hospital costs: The most expensive conditions by payer, 2011," *Statistical Brief 160*, 2009.

[5] J. Russel, "The current management of septic shock," *Minerva medica*, vol. 99, no. 5, pp. 431–458, 2008.

[6] C. S. Calfee, "Precision medicine: An opportunity to improve outcomes of patients with sepsis," *American journal of respiratory and critical care medicine*, no. ja, 2016.

[7] I. Stojkovic, M. F. Ghalwash, X. H. Cao, and Z. Obradovic, "Effectiveness of Multiple Blood-Cleansing Interventions in Sepsis, Characterized in Rats," *Scientific Reports*, vol. 6, no. 24719, pp. 1–11, 2016.

[8] J. Day, J. Rubin, and G. Clermont, "Using nonlinear model predictive control to find optimal therapeutic strategies to modulate inflammation," *Mathematical Biosciences and Engineering*, vol. 7, no. 4, pp. 739–763, 2010.

[9] V. Radosavljevic, K. Ristovski, and Z. Obradovic, "A data-driven acute inflammation therapy," *BMC medical genomics*, vol. 6, no. Suppl 3, p. S7, 2013.

[10] M. Ghalwash and Z. Obradovic, "A data driven model for optimizing therapy duration for septic patients," in *SIAM SDM14 3rd Workshop on Data Mining for Medicine and Healthcare*, 2014.

[11] S. Song, J. Hogg, Z. Peng, R. Parker, J. Kellum, J. A. Papin, *et al.*, "Ensemble models of neutrophil trafficking in severe sepsis," *PLoS Comput Biol*, vol. 8, no. 3, 2012.

[12] V. Radosavljevic, S. Vucetic, and Z. Obradovic, "Continuous Conditional Random Fields for Regression in Remote Sensing." in *Proceedings of the 19th European Conference on Artificial Intelligence (ECAI 2010) Lisbon, Portugal*, 2010.

[13] V. Radosavljevic, K. Ristovski, and Z. Obradovic, "Gaussian Conditional Random Fields for Modeling Patients Response to Acute Inflammation Treatment," in *ICML 2013 workshop on Machine Learning for System Identification*, 2013.

[14] D. Gligorijevic, J. Stojanovic, and Z. Obradovic, "Improving confidence while predicting trends in temporal disease networks," in *SIAM SDM 4th Workshop on Data Mining for Medicine and Healthcare*, 2015.

[15] X. H. Cao, I. Stojkovic, and Z. Obradovic, "Predicting sepsis severity from limited temporal observations," in *International Conference on Discovery Science*, 2014, pp. 37–48.

[16] J. Stojanovic, D. Gligorijevic, and Z. Obradovic, "Modeling customer engagement from partial observations," in *25th ACM International Conference on Information and Knowledge Management (CIKM 2016)*, 2016.

[17] D. Gligorijevic, J. Stojanovic, and Z. Obradovic, "Uncertainty Propagation in Long-term Structured Regression on Evolving Networks," in *Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)*, 2016.

[18] X. H. Cao, I. Stojkovic, and Z. Obradovic, "A robust data scaling algorithm to improve classification accuracies in biomedical data," *BMC bioinformatics*, vol. 17, no. 359, pp. 1–10, 2016.

[19] M. F. Ghalwash, X. H. Cao, I. Stojkovic, and Z. Obradovic, "Structured feature selection using coordinate descent optimization," *BMC Bioinformatics*, vol. 17, pp. 1–14, 2016.

[20] D. Gligorijevic, J. Stojanovic, and Z. Obradovic, "Disease types discovery from a large database of inpatient records: A sepsis study," *Methods*, vol. 111, pp. 45–55, 2016.

[21] D. Gligorijevic, J. Stojanovic, N. Djuric, V. Radosavljevic, M. Grbovic, R. J. Kulathinal, and Z. Obradovic, "Large-scale discovery of disease-disease and disease-gene associations," *Scientific Reports*, vol. 6, 2016.

[22] J. A. Kellum, M. Song, and R. Venkataraman, "Hemoadsorption removes tumor necrosis factor, interleukin-6, and interleukin-10, reduces nuclear factor-κb dna binding, and improves short-term survival in lethal endotoxemia*," *Critical care medicine*, vol. 32, no. 3, pp. 801–805, 2004.

[23] M. Marcellino, J. H. Stock, and M. W. Watson, "A comparison of direct and iterated multistep ar methods for forecasting macroeconomic time series," *Journal of econometrics*, vol. 135, no. 1, pp. 499–526, 2006.

[24] C. E. Rasmussen, *Gaussian processes for machine learning*. MIT Press, 2006.

[25] I. Stojkovic, V. Jelisavcic, V. Milutinovic, and Z. Obradovic, "Distance based modeling of interactions in structured regression," in *Procedeengs of the 25th International Joint Conference on Artificial Intelligence (IJCAI-16)*, 2016, pp. 2032–2038.