MEDINFO 2019: Health and Wellbeing e-Networks for All
L. Ohno-Machado and B. Séroussi (Eds.)

doi:10.3233/SHTI190235

Improving Mechanical Ventilator Clinical Decision Support Systems with a Machine Learning Classifier for Determining Ventilator Mode

Gregory B. Rehm[a], Brooks T. Kuhn[b], Jimmy Nguyen[b], Nicholas R. Anderson[b], Chen-Nee Chuah[a], Jason Y. Adams[b]

[a] University of California Davis, Davis CA 95616, USA
[b] University of California Davis Medical Center, Sacramento CA 95817, USA

Abstract

Clinical decision support systems (CDSS) will play increasing role in improving quality of medical care for critically ill patients. However, due to limitations in current informatics infrastructure, CDSS do not always have complete information on state of supporting physiologic monitoring devices, which can limit input data available to CDSS. This is especially true in use case of mechanical ventilation (MV), where current CDSS have no knowledge of critical ventilation settings, such as ventilation mode. To enable MV CDSS make accurate recommendations related to ventilator mode, we developed a highly performant machine learning model that is able to perform per-breath classification of five of most widely used ventilation modes in USA with average F1-score of 97.52%. We also show how our approach makes methodologic improvements over previous work and is highly robust to missing data caused by software/sensor error.

Keywords:

Artificial respiration, clinical decision support systems, machine learning

Introduction

Mechanical ventilation (MV) is life-saving intervention delivered in intensive care unit (ICU) to patients with acute respiratory failure. When delivered properly, MV allow injured lungs heal while ventilator performs majority of work of breathing for patient. When delivered improperly, MV has been associated with variety of adverse clinical outcomes including patient discomfort, increased sedative dosing, longer ICU length of stay, increased chance of ventilator-induced lung injury, and lower survival [1,2]. New generation of clinical decision support systems (CDSS) promises to reduce chances of delivering improper MV by automating aspects of ventilator configuration, and providing clinically accurate and relevant alerts to providers. However, key detriment to these systems is lack of access to configured state of ventilator and therefore lack information that may improve efficiency of these CDSS [3].

One such piece of information that many MV CDSS lack is choice of ventilation mode (VM) that determines pattern of flow and pressure delivery with each breath (Figure 1 B-D). This information is generally unavailable to CDSS due to lack of interoperability and information exchange between CDSS and ventilator or electronic health record [3]. CDSS knowledge of VM is important because changing VMs may be a necessary procedure in course of patient care [4]. For example, if CDSS determines that patient is breathing asynchronously with ventilator, it may be able to make recommendation that providers choose a different VM that provides more comfort and flexibility in breathing to patients [5-8]. Another example would be that CDSS could provide alerts to clinicians if patients continually violate safe volumes of air to inhale. This would be especially important in cases where patients have acute respiratory distress syndrome and need limited tidal volumes [9,10]. In this case CDSS could recommend patients be placed on VM that limits tidal volumes such as volume-control.

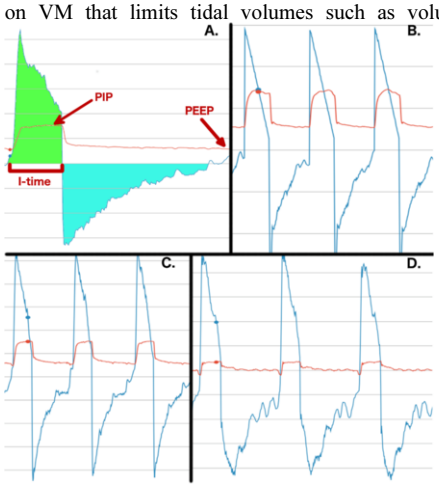

Figure 1- Displays visualizations of ventilator waveform data (VWD). Flow measurements represented in blue, and pressure in red. A.) Here we display examples of how to extract information from VWD. Positive End Expiratory Pressure (PEEP) is noted as minimum pressure supplied by ventilator, and peak inspiratory pressure (PIP) is maximum pressure supplied during inhalation. Inspiratory time (I-time) is amount of time patient breathes in. Total amount of air breathed in represented in green, and air breathed out shown in teal. B.) Shows canonical example of volume control (VC), mode patient receives fixed volume of air for each breath. C.) Shows example of pressure control (PC). In PC, pressure is fixed during inhalation. D.) Example of continuous positive airway pressure (CPAP). Here minimal pressure support is given, and all breaths initiated by patient.

If MV CDSS lacks knowledge of VM from more traditional methods, it may still be able to access it by utilizing information derived from streams of flow and pressure readings that comprise ventilator waveform data (VWD). To the best of our knowledge, only one previous effort has developed rule-based classifier using analysis of VWD for providing hourly VM classifications. However, its use of closed dataset, limited temporal resolution, and accuracy of model represent potential limitations both for research and decision support [3,5]. Having

highly granular temporal resolution VM classification results is important because in practice providers may change VM frequently based on changes in clinical state or patient tolerance of VM. These changes may cause specific VM to remain constant for as low as minutes of time. To improve upon previous work, we note that machine learning (ML) has proven capable of accounting for highly variable nature of physiologic data such as VWD on temporally granular time scales [11,12]. So we created a ML model that could identify different VMs on per-breath basis, with freely accessible dataset, using only VWD as input.

In this paper, we detail multiple important considerations for modeling ML classifier that can classify VM. First, we discuss how we created one of the largest datasets of per-breath labeled information, extraction of features from VWD, and performance of our resulting ML model that can determine five of most widely used ventilation modes in USA [4]. Second, we discuss experiments of how well our model performs in presence of missing training data. Finally, we discuss experimentation we conducted for reducing size of our training dataset by nearly 72% while maintaining generalizability of our classifier to our testing set. To allow reproducibility of our work, our code and dataset are publicly accessible and published on GitHub. Thus, we hope that our work will serve as catalyst for continuing to improve capabilities and efficiency of MV CDSS.

Methods

In this study, we used dataset of VWD collected from 103 subjects (IRB# 647002) within intensive care environments of University of California Davis Medical Center (UCDMC) consisting of MV flow and pressure measurements sampled at 50 Hz [13,14]. Ventilation mode was not recorded in course of VWD data collection. We then randomly selected 2-4 hour epochs of VWD from the 103 subjects. All VWD was stored in data files of 2 hours in length, and approximately 2,000 breaths were stored per data file. Each breath in these epochs was annotated by three expert clinicians (JYA, BTK, JN) for presence of one of five VMs: volume control (VC), pressure control (PC), pressure support (PS), continuous positive airway pressure (CPAP), and proportional assist ventilation (PAV) (Table 1). Many patients had 2-4 hour periods selected where VM was switched multiple times, other modes such as pressure regulated volume control (PRVC), volume support, and airway pressure release ventilation (APRV) were found, and annotated within these epochs, but were excluded in our final analysis because of their rarity of use at UCDMC.

Table 1-Descriptive statistics for our dataset for each ventilator mode analyzed. Also analyzed number of patient ventilator asynchrony (PVA), suction, and cough breaths found [14]. While these breaths do not represent normal breathing, they are typical in clinical practice.

	Volume Control	Pressure Control	Pressure Support	CPAP	PAV
Patients	23	37	55	28	22
Total Breaths	61,662	78,635	92,360	14,795	36,303
PVA Breaths	7,714	4,570	6,924	2,373	7,669
Suction Breaths	750	136	681	350	373
Cough Breaths	229	117	178	56	96

Given VWD is so heterogeneous it can be difficult for even expert clinicians to make consistent classification of breathing patterns [15]. Thus, in performing classification of VM we ensured that each breath was dual clinician adjudicated, meaning that two clinicians would independently annotate a single breath, and if classifications disagreed they would be resolved through communication between the two [14]. To further account for breathing heterogeneity, we included regions containing pathologic patient-ventilator interactions such as patient-ventilator asynchrony (PVA), routine clinical events such as suctioning and cough, and regions of noisy data caused by moisture/blood/mucus in ventilation circuit tubing [14].

Table 2-Set of proposed features for our model. Features were segmented into per-breath and multi-breath time frames.

Feature	Description
Inspiratory Flow Slope Variance (per breath)	This feature measures variance of successive, 0.08-second long slope measurements of inspiratory flow curve of a single breath. This feature was effective for classifying volume control.
Variance of Pressure (per breath)	This feature takes variance of all pressure measurements for a single breath. This feature was helpful for classifying CPAP which typically utilizes low pressures relative to PEEP on inspiration.
Variance of Per-Breath Inspiratory Flow Slope Variance	The inspiratory flow slope variance was found on per breath basis, and this feature takes variance of inspiratory flow slope variance across a 10 breath window.
Inspiratory Time (I-time) Variance (10 breath window)	The amount of time that patient inhales for single breath is called I-time. This feature calculated variance of 10 successive breaths.
Pressure-Based I-time Variance (10 breath window)	We defined pressure-based I-time as amount of time (seconds) that pressure is elevated by [0.4 * (PIP - PEEP)] above PEEP. This was an important variable to measure in pressure control and pressure support, where flow-based I-time can be shorter than ventilator's set I-time, which may occur in delayed cycling asynchrony.
N Plateau Pressures (20 breath window)	A plateau pressure is taken on ventilator when inspiratory flow is set to 0 for a certain amount of time, during which ventilator can read residual pressure in respiratory system. PAV will repetitively take plateau pressures in order to adjust ventilation to patient's needs.
Pressure-Based I-time Variance (100 breath window)	In this feature, pressure-based I-time statistic is also calculated for 100-breath window. This feature was necessary to provide capacity for differentiating between pressure control and pressure support in synchronously breathing patients.

With this dataset, we utilized 55 patients and 140,928 breaths for our training cohort, and 48 patients and 165,988 breaths for our testing cohort. There was no patient overlap between testing and training cohorts. Testing set was chosen to be approximately as large as training set because initial modeling

yielded strong results, and we wished to utilize large testing set as further validation for our approach. Using both Scikit-learn and Pytorch ML libraries [16,17], we evaluated use of multiple ML algorithms including: support vector machine (SVM) [18], multi-layer perceptron (MLP), long-short term memory recurrent neural network (LSTM RNN) [19], logistic regression, and random forest (RF) classifier [20]. All models performed classification on per-breath basis, highest possible level of granularity possible in VM classification. Based on model investigation, we settled on usage of RF with parameterization of 30 classifier trees for our final model (see online supplemental).

Our feature set is composed of 7 items of expert-guided information derived from raw VWD, and is described in Table 2. Our features are derived from both per breath and multi-breath analytic time frames. Per-breath time frames occur over single breath, while multi-breath time frames are composed of windows of short, medium, and long periods of breathing. Short window is 10 breaths long, medium window 20 breaths, and long window 100 breaths. Tuning of features and hyperparameters was guided by performing 10-fold cross-validation of our training data. After tuning model hyperparameters during the validation phase, we evaluated our model on our testing set. No additional changes to our feature set, or model hyperparameters were performed after model development was completed in training set. Model performance is primarily reported through F1-score because it is more representative of class-imbalanced classifier performance than accuracy is. F1-score is calculated as harmonic mean of precision (PPV) and recall (sensitivity):

$$\text{F1-score} = 2\frac{precision * recall}{precision + recall}$$

Limitation to using RF to classify ventilator mode is RF classifier assumes that all breaths are independent of each other. However, ventilator mode is a continuous setting that does not vary over time, unless it is manually changed by provider. Therefore, one breath's mode is often predictive of next breath's mode. This modeling incongruity causes RF classifier to sometimes perform incorrect VM classification even in periods where classifier correctly predicts correct VM for a majority of breaths. To smooth these incorrect predictions, we implement an algorithm we term "look-ahead smoothing" which operates as second pass heuristic on all per breath RF breath predictions. More specifically, once RF is finished, look-ahead smoothing examines each breath VM classification sequentially, and if it determines breath's classification is not in accordance with previous n breaths then it will look ahead at next n breaths in sequence. The breath will then be re-classified in accordance to majority x percent of subsequent n breaths. Both n and x are configurable parameters that we set at $n = 50$ and $x = 60$, parameters which were found via sensitivity analysis. In real-time classification, assuming average respiratory rate of 20 breaths per minute, this technique results in latency of at most 2.5 minutes between a given breath and availability of its final classification.

Finally, we implemented experiment to test how well our classifier would generalize to larger dataset if random breaths in our training dataset were missing due to some technical error. So we conduct experiment where we ablate (i.e. remove) data observations at random from our training dataset in equal proportion for VC, PC, PS, CPAP, and PAV. We do not perform any ablation on the testing set. We then report results of this experiment by recording F1-score for each class with respect to percentage of dataset that simulated as missing.

Results

Using RF model with feature set defined in Table 2, we initially performed 10-fold cross validation with our training set to test performance of our VM classifier. We found that during cross validation our model consistently performed within 98-99% for F1-score, recall, and specificity for all VMs. We then evaluated our model on withheld test set. CPAP suffered largest drop in performance because it confused PS for CPAP for an entire patient. VC/PAV suffered no drop in performance and PC/PS only suffered slight declines in performance (Table 3).

Table 3-Performance of our Random Forest model when applied to our withheld testing set.

Mode	F1-Score	Accuracy	Precision	Recall	Specificity
VC	0.999	1.0	0.998	1.0	1.0
PC	0.989	0.993	0.983	0.996	0.992
PS	0.975	0.981	0.993	0.958	0.996
CPAP	0.85	0.988	0.767	0.952	0.989
PAV	0.994	0.999	0.99	0.998	0.999

We hypothesized that since the model performed well on both training and testing sets that it would also be robust to scenarios in which breath data went missing due to reason of sensor or software failure. We report results for this experiment in Figure 2. We found model is robust to missing data until approximately 90% of data is removed. After this point PC and PS F1-score performance begins to decrease and other classifications begin to fluctuate. After 99% of data is removed our classifications lose clinical utility.

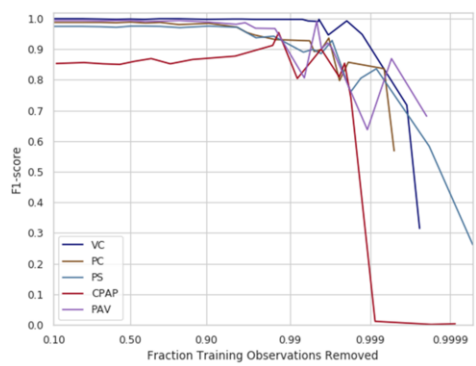

Figure 2-Here we simulate scenario where percentage of training observations is missing due to some kind of software/hardware error.

Given results of random ablation experiment, we hypothesized that we may have created too large a training set. To reduce the size of our training set in generalizable, non-random way, we hypothesized we only needed to keep the first of certain number of contiguous breaths from each VM per data file, and still maintain performance of our original model. In this respect, we could make recommendations to physicians to only annotate first $m$breaths in a series and just leave the rest alone. This could also decrease amount of time necessary to annotate VM on future patients. So, we performed a sensitivity analysis to determine what the optimal number of contiguous observations to keep per ventilator mode is. We do this by sequentially iterating over each VM in our training set and only picking first m breaths in a file while keeping number of observations from other VMs constant. Our analysis (Figure 3) showed that it was

most optimal to only use first 450 VC observations, first 120 PC, 1,200 PS, 160 CPAP, and 80 PAV observations in a file. Using this methodology, we ablated overall number of training observations by 71.41% from 140,928 to 40,285 observations, while still maintaining generalizability of our training set to our withheld test set, and largely improved CPAP performance (Table 4). By performing ablation we were able to boost average F1-score of our classifier to 0.9752 from 0.9614 that was reported in Table 3.

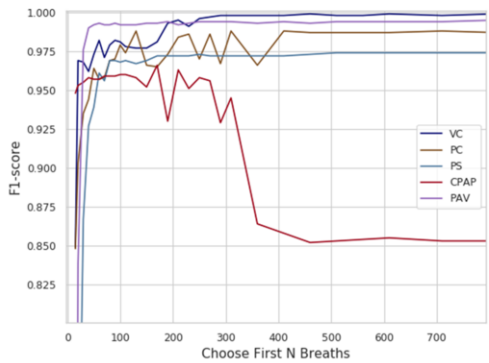

Figure 3-Results from our sensitivity analysis for choosing first N contiguous breaths for a given mode in a data file.

Table 4-Final results of our ablation experiment where we only keep first 450 VC, 120 PC, 1,200 PS, 160 CPAP, and 80 PAV observations in a data file. We note final number of training observations that we kept, and report how much of reduction that was in contrast to original training set. Performance improvements/degradation over results listed in Table 3 are bracketed alongside final performance metrics. E.g. performance increase of 2.0% is denoted as (+.02).

Mode	Training Observations	F1-Score
VC	6,079 (-83.65%)	0.998 (-0.001)
PC	2,154 (-92.77%)	0.964 (-0.025)
PS	27,892 (-26.81%)	0.967 (-0.008)
CPAP	3,040 (-73.55%)	0.955 (+0.105)
PAV	1,120 (-94.46%)	0.993 (-0.001)

Discussion

In this paper, we highlighted how we created dataset of 308,957 breaths annotated for VM on per-breath basis and how we developed highly accurate, ML-based VM classification model that only utilizes raw VWD to perform classifications. Our VM classifier was highly performant in detecting five of most widely used VMs in USA, even in presence of signal noise, episodes of PVA, and routine clinical events such as cough and suction [4]. Using our approach, we were able to achieve methodological and performance improvements in VM classification compared to current state of art [3]. In this regard, Murias reported 89% accuracy at detecting per-hour VM classification, and we report average accuracy of 98.05% of per-breath VM classification (Table 3). Finally, we examined how robust our model is to presence of missing training data, and additional experimental results that suggested how we can decrease the size of our dataset while still maintaining generalizability of our classifier.

We took multiple measures to ensure we were not overtraining our classifier. First, we utilized a relatively large and diverse sampling of patients to create both our testing and training sets. This created one of the largest available datasets of per-breath labeled VWD. Two to four hour epochs were chosen at random from each of these patients. Our testing set included almost as many patients as our training set, and was composed of more breaths than our training set. There was no overlap of patients between training and testing sets. Finally, all model features and hyperparameters were evaluated on the training set using K-fold validation, and were frozen after initial evaluation of our testing set.

Our ablation experiments deserve additional consideration. Results of the random ablation experiments highlight multiple things: 1) RF is extremely performant with our featurization approach, and is also able to perform VM classification with small amounts of data. 2) Our ablation results also illustrate that it may not be necessary to create very large training datasets of information to create performant ML classifiers for VM. 3) Our size reduction experiments did see some decreases in performance in PC and PS because of the manner in which we performed our sensitivity analysis. In our analysis we only modified observations from a single VM type while keeping other VM observations constant, so it was not possible to determine side effects from simultaneously ablating several modes at once. Future experiments could perform more computationally demanding task of ablating multiple modes at once to further explore the issue. 4) Our size reduction experiment showed that first 160 breaths seem to be most representative of CPAP breathing patterns. We hypothesize this can be explained by the fact that some patients tire quickly when on CPAP, and thus their breathing can become more irregular. In this case, later breaths in CPAP sequences may more closely resemble asynchronous breathing from other ventilator modes instead of CPAP.

This work had several limitations. Our use of "look-ahead smoothing" introduced small latency of 2.5 minutes to real-time ventilator mode predictions. This time delay in classification would not likely be of clinical consequence since CDSS recommendations requiring VM state information would rarely be executed over such short time frames to ensure that transient changes in waveforms do not trigger frequent false alarms. If latency is not desired then "look-behind smoothing" can be used as alternative approach. Our study was also confined to a single academic medical center and single ventilator type. There are also additional ventilator modes such as PRVC that we were unable to add to our model due to their paucity of use at UCDMC. We welcome additional collaboration and inclusion of multi-center data and have publicly released our code and dataset.

Conclusions

In conclusion, we created a highly-performant ML classifier for detecting five of most commonly used ventilator modes in USA, using only raw VWD as input. Our use case further demonstrates utility of ML analysis of physiologic waveform data to improve CDSS characterization of patient state when state is missing due to limitations of available informatics infrastructure. We also illustrated usage of dataset ablation to characterize how missing data affects generalization performance of our classifier, and how we can restrict size of our training set while maintaining model generalization to our test dataset. Our classifier will facilitate development of more advanced automated MV CDSS to improve management of patients experiencing respiratory failure.

Acknowledgements

This work was generously supported by National Heart, Lung, and Blood Institute (NHLBI) Individual Predoctoral Fellowship (award number 1F31HL144028-01), and NHLBI Emergency Medicine K12 Clinical Research Training Program (grant number K12 HL108964).

References

[1] A. Slutsky, and M. Ranieri, Ventilator-induced lung injury, *N Engl J Med* **369** (2013), 2126-2136.
[2] A. Bagchi, M.I. Rudolph, P.Y. Ng, F.P. Timm, D.R. Long, S. Shaefi, K. Ladha, M.F. Vidal Melo, and M. Eikermann, The association of postoperative pulmonary complications in 109,360 patients with pressure-controlled or volume-controlled ventilation, *Anaesthesia* **72** (2017), 1334-1343.
[3] G. Murias, J. Montanyà, E. Chacón, A. Estruga, C. Subirà, R. Fernández, B. Sales, C. de Haro, J. Lopez-Aguilar, U. Lucangelo, J. Villar, R.M. Kacmarek, and L. Blanch, Automatic detection of ventilatory modes during invasive mechanical ventilation, *Crit Care* **20** (2016), 258.
[4] T. Pham, L.J. Brochard, and A.S. Slutsky, Mechanical ventilation: state of the art, *Mayo Clin Proc* **92** (2017), 1382-1400.
[5] N. Rittayamai, C.M. Katsios, F. Beloncle, J.O. Friedrich, J. Mancebo, and L. Brochard, Pressure-controlled vs volume-controlled ventilation in acute respiratory failure: a physiology-based narrative and systematic review, *Chest* **148** (2015), 340-355.
[6] G. Murias, U. Lucangelo, and L. Blanch, Patient-ventilator asynchrony, *Curr Opin Crit Care* **22** (2016), 53-59.
[7] K.G. Mellott, M.J. Grap, C.L. Munro, C.N. Sessler, P.A. Wetzel, J.O. Nilsestuen, and J.M. Ketchum, Patient ventilator asynchrony in critically ill adults: frequency and types, *Heart Lung* **43** (2014), 231-243.
[8] D.L. Grieco, M.M. Bitondo, H. Aguirre-Bermeo, S. Italiano, F.A. Idone, A. Moccaldo, M.T. Santantonio, D. Eleuteri, M. Antonelli, J. Mancebo, and S.M. Maggiore, Patient-ventilator interaction with conventional and automated management of pressure support during difficult weaning from mechanical ventilation, *J Crit Care* **48** (2018), 203-210.
[9] Acute Respiratory Distress Syndrome Network, R.G. Brower, M.A. Matthay, A. Morris, D. Schoenfeld, B.T. Thompson, and A. Wheeler, Ventilation with lower tidal volumes as compared with traditional tidal volumes for acute lung injury and the acute respiratory distress syndrome, *N. Engl J Med* **342** (2000), 1301-1308.
[10] L. Gattinoni, J.J. Marini, A. Pesenti, M. Quintel, J. Mancebo, and L. Brochard, The 'baby lung' became an adult, *Intensive Care Med* **42** (2016), 663-673.
[11] G. Rehm, J. Han, B. Kuhn, J.P. Delplanque, N. Anderson, J. Adams, and C.N. Chuah, Creation of a robust and generalizable machine learning classifier for patient ventilator asynchrony, *Methods Inf Med*, **57** (2018), 208-219.
[12] P.D. Sottile, D. Albers, C. Higgins, J. Mckeehan, and M.M. Moss, The association between ventilator dyssynchrony, delivered tidal volume, and sedation using a novel automated ventilator dyssynchrony detection algorithm, *Crit Care Med* **46** 2018, e151-e157.
[13] G.B. Rehm, B.T. Kuhn, J.P. Delplanque, E.C. Guo, M.K. Lieng, J. Nguyen, N.R. Anderson, and J.Y. Adams, Development of a research-oriented system for collecting mechanical ventilator waveform data, *J Am Med Inform Asso*, **25(** 2017), 295-299.
[14] J.Y. Adams, M.K. Lieng, B.T. Kuhn, G.B. Rehm, E.C. Guo, S.L. Taylor, J.P. Delplanque, and N.R. Anderson, Development and validation of a multi-algorithm analytic platform to detect off-target mechanical ventilation, *Sci Rep* **7** (2017), 14980.
[15] D. Colombo, G. Cammarota, M. Alemani, L. Carenzo, F.L. Barra, R. Vaschetto, A.S. Slutsky, F. Della Corte, and P. Navalesi, Efficacy of ventilator waveforms observation in detecting patient-ventilator asynchrony, *Crit Care Med* **39** (2011), 2452-2457.
[16] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, Scikit-learn: machine learning in python, *J Machine Learn Res* 12 (2011), 2825-2830.
[17] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, Automatic differentiation in pytorch, In: *31st Conference on Neural Information Processing Systems*, Long Beach, CA, USA, 2017, 1-4.
[18] M.A. Hearst, S.T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, Support vector machines, *IEEE Intelligent Systems* **13** (1998), 18-28.
[19] F.A. Gers, J. Schmidhuber, and F. Cummins, Learning to forget: continual prediction with LSTM, In: *9th International Conference on Artificial Neural Networks: ICANN '99*, Edinburgh, UK, 1999, 850-855.
[20] L. Breiman, Random Forests, *Machine Learning*, **45** (2001), 5-32.

Address for correspondence

For correspondence: Gregory B. Rehm: grehm@ucdavis.edu