Machine Learning Prediction of Blood Potassium at Different Time Cutoffs

Jake A. Bergquist[1,2,3], Deekshith Dade[1,5], Brian Zenger[6], Rob S. MacLeod[1,2,3], Xingyang Ye[4], Ravi Ranjan[2,3,4], Tolga Tasdizen[1,5], Benjamin A. Steinberg[4,7]

[1] Scientific Computing and Imaging Institute, University of Utah, SLC, UT, USA
[2] Nora Eccles Treadwell CVRTI, University of Utah, SLC, UT, USA
[3] Department of Biomedical Engineering, University of Utah, SLC, UT, USA
[4] School of Medicine, University of Utah, SLC, UT, USA
[5] Department of Electrical and Computer Engineering, University of Utah, SLC, UT, USA
[6] Washington University School of Medicine, Saint Louis, MO, USA
[7] University of Colorado Anschutz Medical Campus, Denver, CO, USA

Abstract

Because serum potassium and ECG morphology changes exhibit a well-understood connection, and the timeline of ECG changes can be relatively quick, there is motivation to explore the sensitivity of ML based prediction of serum potassium using 12 lead ECG data with respect to the time between the ECG and potassium readings.

We trained a convolutional neural network to classify abnormal (serum potassium above 5 mEq/L) vs normal (serum potassium between 4 and 5 mEq/L) from the ECG alone. We compared training with ECGs and potassium measurements filtered to be within 1 hour, 30 minutes, and 15 minutes of each other. We explored scenarios that both leveraged all available data at each time cutoff as well as restricted data to match training set sizes across the time cutoffs. For each case, we trained five separate instances of our neural network to account for variability.

The 1 hour cutoff with all data resulted in an average area under the receiver operator curve (AUC) of 0.850 and a weighted accuracy of 76.3%, 15 minutes resulted in 0.814, 72.5%, and 30 minutes. Truncating the training sets to the same size as the 15 minute cutoff results in comparable average accuracy and AUC for all. Our future studies will continue to explore the performance of ML potassium predictions through investigations of failure cases, identification of biases, and explainability analyses.

1. Introduction

Advances in machine learning electrocardiogram (ECG) analysis have led to a variety of clinical and research tools that can rapidly identify disease states using 12-lead ECG signals alone.[1, 2] Such tools are often able to both provide superior performance over traditional ECG analysis methods as well as identify conditions which cannot be identified using traditional analyses.[1, 2] Such advances in ECG ML tools have included prediction of blood serum potassium levels, enabling rapid, sensitive, and automated monitoring that would otherwise require blood tests.[3, 4]

Serum potassium and ECG morphology changes exhibit a physiologically well-known connection that can be highly time dependant. Because of this, exploration of ML-ECG analysis tasks in the context of serum potassium present several unique opportunities. For example, because serum potassium levels cause well understood ECG changes, ML-ECG analysis of serum potassium may provide more transparent explainability and interpretability for how the ML tools are assessing the ECGs. The temporal relationship between serum potassium changes and ECG changes also presents an interesting avenue of exploration. In the past ML tools have proven able to predict patient characteristics and disease states not thought possible with the ECG alone, and so it may be that such tools could also resolve more difficult tasks in the realm of ECG-potassium changes such as predicting future potassium fluctuations using temporally distant ECG signals. As a first step in exploring these various avenues, we were motivated to interrogate the performance of ML potassium classification tasks under different times between the ECG and potassium readings.

Previous studies in application of ML to serum potassium classification have focused on specific datasets of homogeneous patients such as Yasin *et al.*,[4] which used a dataset of patients undergoing repeated potassium measurements due to kidney dialysis. This and other studies[3] have also focused on using limited or single lead ECGs either due to limited availability of 12-lead datasets or due to interest in pursuing ML tools that can work on single

Table 1. Dataset subset summary. Each subset is defined by the time between ECG and potassium value as well as the training set size. Training set balance is shown as a percent of the total training dataset size.

Time Cutoff (seconds)	Test Size	Train Size	Training Normal : Abnormal
900	7,014	77,124	91% : 9%
1800	11,492	129,052	92% : 8%
1800	11,492	78,078	92% : 8%
3600	16,945	184,717	92% : 8%
3600	16,945	79,321	92% : 8%

lead devices. In the present study, we leverage a dataset of paired 12-lead clinical ECGs and serum potassium measurements from a more diverse cohort which includes all patients in the University of Utah health system who have at least one potassium test an one ECG. We split our initial cohort into three groups based on time between ECG and potassium measurement: 15 minutes (900 seconds), 30 minutes (1,800 seconds), and one hour (3,600 seconds).

2. Methods

Dataset: Digital ECG recordings were collected from 444,026 University of Utah Health patients from 2012 to 2021. Each ECG measured consisted of 8 leads(L1, L2, V1 through V6) and between 5 and 10 seconds of continuous simultaneous recording from each lead at 500 Hz. Matched patient specific serum potassium lab values were extracted from the University of Utah health database with the help of the electronic data warehouse service at the University of Utah. Patients under the age of 18 or over the age of 90 were excluded. All studies and data acquisition were subject to and complied with University of Utah institutional review board review and requirements.

Within this dataset, we identified 42,440,729 ECG to blood serum potassium measurement pairs. We then truncated this larger dataset to include only normal or abnormally high potassium values (4 mEq/L to 5 mEq/L: normal, above 5 mEq/L: abnormal), and subset this according to three time cutoffs between ECG and potassium test: 15 minutes (900 seconds), 30 minutes (1,800 seconds), and one hour (3,600 seconds). For each time cutoff we randomly divided the data into a 90% training and 10% testing subset. To control for varied training set sizes, we also created a truncated training set for the 15 minute and 1 hour subsets, reducing the training cohort size to be near the size of the 15 minute subset. Table 1 summarizes the datasets.

Machine Learning Architecture and Training: We formulated the detection of abnormal serum potassium framed as a binary classification task using 4 mEq/L to 5 mEq/L as normal potassium ranges, and above 5 mEq/L as abnormal. Our network architecture is based on a residual network that we have shown to be an effective structure for ML-ECG analysis [5, 6]. In brief, the network consisted of temporal and spatial convolutional filters, batch normalization, dropout (probability = 0.5), a rectified linear unit (ReLU), fully connected layers, and a sigmoid output. Spatial and temporal convolutional layers were arranged into residual blocks, and their output features were concatenated before the fully connected layers. The architecture is depicted in Figure 1. Input data consisted of the 8 leads of the ECG stacked along one dimension and time along the other. All ECG recordings above 5 seconds in length (2,5000 samples) were randomly cropped to 2,500 samples in the time dimension.

For each dataset described in Table 1, 5 replicates of training were performed on separately randomly initialized networks. Each replicate used the same train and test split, but with different batch randomization in training. Training consisted of 50 epochs using the ADAM optimizer. At each epoch, area under the receiver operator curve (AUC) was computed in the training and testing datasets.

Analysis metrics: For each network we computed AUC and class weighted accuracy using the network parameters from the epoch with the highest test dataset AUC.

3. Results

Table 1 summarizes the results. Truncating the training sets to the same size as the 15 minute cutoff results in comparable accuracy and area under the receiver operator curve (AUC) for all time cutoffs. The highest performance was seen in the 1 hour (3,600 second) non-truncated dataset, which achieved an average weighted accuracy of 76.3% and average AUC of 0.85. Variability in AUC values was low, never exceeding 0.0082. However, variability in weighted accuracy was more heterogeneous, and did not follow a predictable trend. The highest variability was seen in the 30 minute (1,800 second) non-truncated dataset at 4.3%, followed by the 1 hour (3,600 minute) truncated dataset at 3.7%. The network performance metrics are shown in Figure 2, where the individual performance of each of the 5 replicates per training scenario can be seen.

4. Discussion and Conclusions

In this study we explored the classification of elevated vs normal serum potassium levels across three time cutoffs between ECG and blood test. We found that, when training set size was controlled for, all time cutoffs produced similar weighted accuracy (around 72%) and AUC scores (around 0.82). We had theorized that the 1 hour prediction task would be the most difficult, however, in this

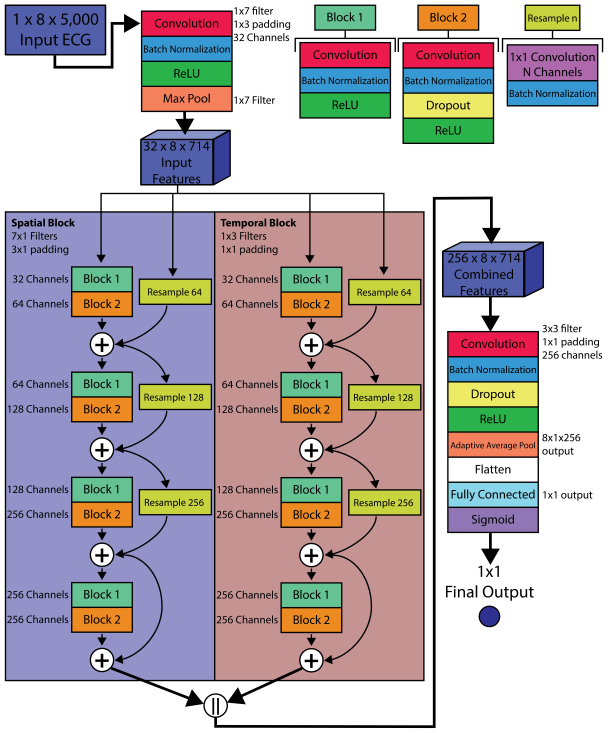

Figure 1. ECG ML architecture. This network consists of an input stage, temporal and spatial residual blocks, and an output stage. Each residual block consists of four layers of residual blocks, similar to the common resnet structure. In cases where the number of input channels is less than the output channels (layers 1 through 3), the input is re-sampled using a 1x1 convolutional layer. The spatial residual block uses 7x1 convolutional filters whereas the temporal uses 1x3 filters. The features from the two residual blocks are concatenated before the output stage. When single-leads are used, the spatial blocks instead use 1x1 convolutional filters.

case we saw the highest performance metrics with an average weighted accuracy of 76.3% and average AUC of 0.85. This may be explained by the substantially larger training dataset available for the 1 hour task (roughly 60,000 more samples than the 30 minute cutoff and roughly 100,000 more samples than the 15 minute cutoff), a result which possibly reinforces the heuristic in ML fields that dataset size is the most important factor. However, even when we controlled for dataset size by reducing each time cutoff dataset to roughly the size of the smallest (around 77,000 training samples), we still observed similar accuracy and AUC scores.

One line of reasoning that may elucidate our findings relates to the different clinical scenarios that are likely at each cutoff. Patients who are receiving multiple serum potassium measurements coupled with short time delays to ECGs are likely patients in an acute care scenario, whereas patients with few potassium measurements and distantly spaced ECGs are likely more stable, as outpatients. The former short term group are likely experiencing more dynamic ECG and potassium dynamics that may be more difficult to predict, while the latter long term group are likely experiencing more stable ECG and potassium dynamics. In future studies we will further stratify our datasets by the clinical scenario (outpatient vs acute) in order to better understand the performance and limitations of these ML techniques in different application cases.

Potassium measure to ECG relationships may also be complicated by the sequence of tests; was the ECG before or after the potassium measurement? Our present study did not discriminate, however future research may benefit from restricting or differentiating between ECG-before and ECG-after cases. Additionally, potassium serum measurements are subject to errors, and this may be partially mitigated by, for example, omitting samples where the specimen was hemolyzed. However, full knowledge of this information in the dataset may be lacking.

It has been suggested in discussions around the performance of ML tools in ECG diagnosis that prediction of any specific disease phenotype (low ejection fraction, high potassium, *etc.,*), is simply prediction of healthy v.s. unhealthy patients generally. In recent research, we sought to identify possible confounding factors associated with ML diagnostic performance,[6] and here we noted that patients with other comorbidities often corresponded to poor ML prediction accuracy. A possible method for addressing these complicating factors would be to exclude patients with abnormal ECG findings, however this may itself introduce both biases and dataset limitations. We propose that a combination of a predictive task in which known ECG phenotype exist (such as high potassium) and ML explainability and visualization tools may lead to elucidation on whether or not the ML algorithm is truly seeing the disease phenotype of interest or simply identify unhealthy patients regardless of disease. We intend to pursue such explainability techniques as a next step in this research.

The present study is limited by the imbalance of the

[T]

Table 2. Network performance in the testing datasets across training scenarios. Weighted accuracy and area under the receiver operator curve (AUC) are shown as ± one standard deviation.

Time Cutoff (seconds)	Train Size	Test Weighted Accuracy	Test AUC
900	77,124	72.5% ±1.1%	0.814 ±0.0032
1800	129,052	72.4% ±4.3%	0.825 ±0.0082
1800	78,078	72.6% ±2.2%	0.823 ±0.0051
3600	184,717	76.3% ±0.5%	0.850 ±0.0030
3600	79,321	72.8% ±3.7%	0.830 ±0.0052

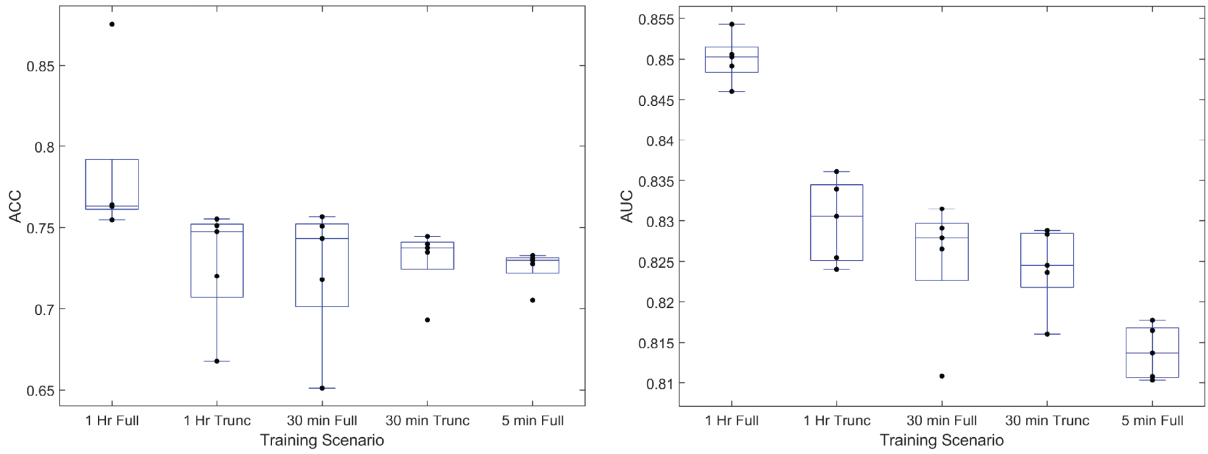

Figure 2. Weighted accuracy (ACC, left) and area under the receiver operator curve (AUC, right) for each network in each training scenario. Training set size is based on Table 1 The 1 hour (3,600 seconds)scenario uses either the full training dataset (184,717 samples) or truncated (79,321 samples). The 30 minute (1,800 seconds) scenario uses either the full training dataset (129,052 samples) or truncated (78,078 samples). The 15 minute (900 seconds) scenario uses the full training dataset of 77,124 samples. The Y axis range in each metric has been reduced to hilight subtle differences between each network and training scenario.

dataset in that most of the potassium measurements were within the normal range. To account for this, we reported weighted accuracy. However, this bias in the training data could lead to poor performance in more balance or abnormal heavy datasets. While class weighting and other training techniques may be used to mitigate this problem, increasing the number of abnormal cases in our dataset will be the best method for improving the robustness of our serum potassium predictions. We are also limited by the time cutoffs chose, ranging from 15 minutes to 1 hour. We chose these as a starting point for our investigation, and plan to expand these to include more acute (less than 15 minutes) and long term (over 1 hour) scenarios.

Acknowledgments

Support for this research came from the Center for Integrative Biomedical Computing (www.sci.utah.edu/cibc), NIH/NIGMS grants P41 GM103545 and R24 GM136986, NIH/NIBIB grant U24EB029012, NIH/NHLBI T32HL007576, K23HL143156, 1R21HL172288-01, University of Utah Data Science Hub, and the Nora Eccles Harrison Foundation for Cardiovascular Research.

References

[1] Jentzer JC, Kashou AH, Attia ZI, Lopez-Jimenez F, Kapa S, Friedman PA, Noseworthy PA. Left ventricular systolic dysfunction identification using artificial intelligence-augmented electrocardiogram in cardiac intensive care unit patients. International Journal of Cardiology 3 2021; 326:114–123. ISSN 1874-1754.

[2] Trayanova NA, Popescu DM, Shade JK. Machine Learning in Arrhythmia and Electrophysiology. Circulation Research 2021;128(4):544–566. ISSN 15244571.

[3] Attia ZI, DeSimone CV, Dillon JJ, Sapir Y, Somers VK, Dugan JL, Bruce CJ, Ackerman MJ, Asirvatham SJ, Striemer BL, Bukartyk J, Scott CG, Bennet KE, Ladewig DJ, Gilles EJ, Sadot D, Geva AB, Friedman PA. Novel bloodless potassium determination using a signal-processed single-lead ecg. Journal of the American Heart Association 2016; 5(1):e002746.

[4] Yasin OZ, Attia Z, Dillon JJ, DeSimone CV, Sapir Y, Dugan J, Somers VK, Ackerman MJ, Asirvatham SJ, Scott CG, Bennet KE, Ladewig DJ, Sadot D, Geva AB, Friedman PA. Noninvasive blood potassium measurement using signal-processed, single-lead ecg acquired from a handheld smartphone. Journal of Electrocardiology 2017;50(5):620–625. ISSN 0022-0736.

[5] Bergquist JA, Zenger B, Brundage J, Shah R, Ye X, Lyons A, MacLeod RS, Ranjan R, Tasdizen T, Bunch TJ, Steinberg BA. Comparison of machine learning detection of low left ventricular ejection fraction using individual ecg leads. In 2023 Computing in Cardiology. 2023; 1–4.

[6] Bergquist JA, Zenger B, Brundage J, MacLeod RS, Bunch TJ, Shah R, Ye X, Lyons A, Torre M, Ranjan R, Tasdizen T, Steinberg BA. Performance of off-the-shelf machine learning architectures and biases in low left ventricular ejection fraction detection. Heart Rhythm O2 2024;5(9):644–654. ISSN 2666-5018.

Address for correspondence:

Jake Bergquist
University of Utah
72 Central Campus Dr, Salt Lake City, UT 84112
jbergquist@sci.utah.edu