2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) | 979-8-3503-2447-1/23/$31.00 ©2023 IEEE | DOI: 10.1109/EMBC40787.2023.10341204

QTNet: Deep Learning for Estimating QT Intervals Using a Single Lead ECG

Ridwan Alam, Aaron D. Aguirre, and Collin M. Stultz, *Members, IEEE*

***Abstract*— QT prolongation often leads to fatal arrhythmia and sudden cardiac death. Antiarrhythmic drugs can increase the risk of QT prolongation and therefore require strict post-administration monitoring and dosage control. Measurement of the QT interval from the 12-lead electrocardiogram (ECG) by a trained expert, in a clinical setting, is the accepted method for tracking QT prolongation. Recent advances in wearable ECG technology, however, raise the possibility of automated out-of-hospital QT tracking. Applications of Deep Learning (DL) - a subfield within Machine Learning - in ECG analysis holds the promise of automation for a variety of classification and regression tasks. In this work, we propose a residual neural network, QTNet, for the regression of QT intervals from a single lead (Lead-I) ECG. QTNet is trained in a supervised manner on a large ECG dataset from a U.S. hospital. We demonstrate the robustness and generalizability of QTNet on four test-sets; one from the same hospital, one from another U.S. hospital, and two public datasets. Over all four datasets, the mean absolute error (MAE) in the estimated QT interval ranges between 9ms and 15.8ms. Pearson correlation coefficients vary between 0.899 and 0.914. By contrast, QT interval estimation on these datasets with a standard method for automated ECG analysis (NeuroKit2) yields MAEs between 22.29ms and 90.79ms, and Pearson correlation coefficients 0.345 and 0.620. These results demonstrate the utility of QTNet across distinct datasets and patient populations, thereby highlighting the potential utility of DL models for ubiquitous QT tracking.**

***Clinical Relevance*— QTNet can be applied to inpatient or ambulatory Lead-I ECG signals to track QT intervals. The method facilitates ambulatory monitoring of patients at risk of QT prolongation.**

I. Introduction

QT interval (QTI) and heart-rate corrected QT interval (QTc) are vital biomarkers for many cardiovascular health conditions and severe adverse outcomes. QT prolongation can result from multiple causes including genetic factors and drug-effects, and can increase the risk for the life-threatening arrhythmias, Torsades-de-Pointes (TdP), and sudden cardiac death. Hence, frequent monitoring of the QT interval is essential for patients diagnosed with a genetic or systemic predisposition to QT prolongation, or for those who need antiarrhythmic medications for other cardiac conditions [1,2]. For example, patients who need to be administered Dofetilide, an antiarrhythmic class-III agent often used to treat atrial fibrillation, require hospital admission just for QT monitoring during drug initiation. They are closely monitored by ECG for 48-72 hours for excessive QT prolongation, and their dosage of medication is regulated accordingly. Such burden, both on the patient and the healthcare system, may be reduced with additional remote monitoring capabilities of QTI and QTc.

QTI measurement by an expert from 12-lead ECG tracings remains the 'gold standard' approach for risk assessment. These measurements are confined to the clinical settings, skill- and labor-intensive, and not suitable for ubiquitous real-time monitoring. Wearable ECG devices, such as smartwatches that capture Lead-I only, have advanced toward achieving reliable ECG quality with agreement of the interval measurements compared to the clinical standard measurements [3,4], encouraging the feasibility of remote monitoring.

Automated analysis of the ECG signal remains a major challenge toward remote QTI monitoring since it is impractical for any expert to annotate continuous ECG streams. Deep learning (DL), machine learning (ML), with deep neural network (DNN) models have shown great promise in ECG analysis for classification tasks, such as detecting atrial fibrillation and other arrhythmias, identifying cardiac abnormalities, LQTS, and others [5-7]. Yet, ML models for regression tasks such as inferring intervals or physiological parameters remain an area of active research [7]. Recent DL models on interval prediction from single- or multilead ECG attempt to circumvent direct regression by formulating the problem as multiclass classification or ECG delineation tasks [8-10].

In this paper, we propose a deep learning ResNet-based regression model, named QTNet, (see Fig. 1), to infer QTI and

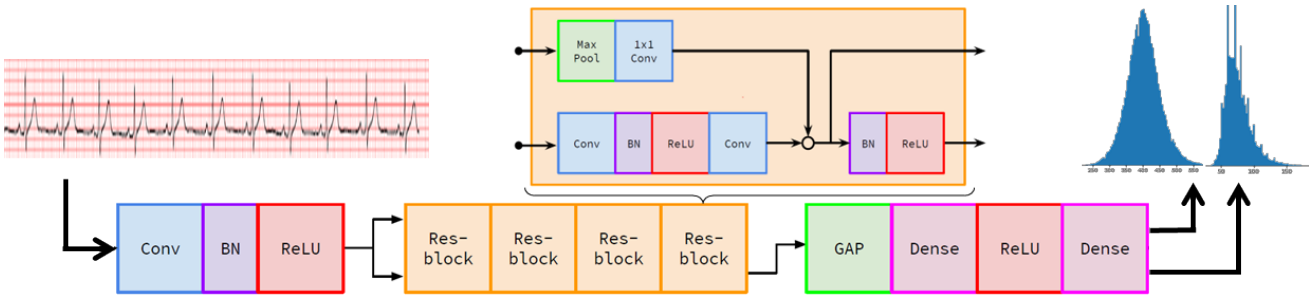

Figure 1. QTNet architecture, adapted from [5]; input: 10-second Lead-I ECG sampled at 250Hz, output: QT interval and heart rate.

R. Alam (corr. author) and C. M. Stultz are with the Massachusetts Institute of Technology, Cambridge, MA 02139, and the Massachusetts General Hospital, Boston, MA 02114. (ridwan@csail.mit.edu, cmstultz@mit.edu).

A. D. Aguirre is with the Wellman Center for Photomedicine, Harvard Medical School, Massachusetts General Hospital, Boston, MA 02114, USA (email: aaguirre1@mgh.harvard.edu).

heart rates (HR) from single channel Lead-I ECG signals. The model is trained and evaluated on more than 4.22 million clinical ECGs from 903.5 thousand patients at the Massachusetts General Hospital (MGH), Boston, MA. The same model is then evaluated, without any fine-tuning, on multiple external datasets to demonstrate generalizability and reliability. External datasets include 3.17 million ECG from the Brigham and Women's Hospital (BWH), Boston, MA, as well as publicly available Physionet datasets on drug-induced QT change, named RDVQ and DMMLD [11-13]. The performance on both internal and external datasets demonstrates low regression errors and high Pearson correlation coefficients. Due to the lack of publicly available Lead-I DL models for QTI regression, we implemented a baseline method using the NeuroKit2 [14] library and compare its performance using the same datasets. QTNet outperforms the baseline in all metrics by notable margins. Hence, the contributions of this work are twofold; first, development of a new DNN model for regression of QT intervals from raw Lead-I ECG, and second, evaluation of the proposed model on external datasets including public resources. The results emphasize the potential of DL in enabling remote QT monitoring applications.

II. Methods

A. Datasets

MGH and BWH contain proprietary ECG databanks (GE MUSE system) for patients who underwent acquisition at any part of their care since 1981. These data contain 12-lead ECG voltages along with metadata including intervals (e.g., QTI) and interpretations (e.g., abnormalities) generated by the ECG devices, often adjudicated by experts. The MGH database is comprised of 4.22M ECG with the QTI and HR labels from 903.5k patients. Similar data exist in the BWH set; 3.17M labelled ECG from 668k patients. The Institutional Review Board of the Mass General Brigham system approved our study to use these data. Though these datasets contain the full 12-lead ECG, only the Lead-I signals are used for model development. All ECG are resampled to 250 Hz sampling rate.

We use only part of the MGH dataset to build QTNet. For training, we use 3.06M ECG (653k patients, 72%), and for validation after each epoch, 534k ECG (115k patients, 13%). The rest 633k ECG (135.5k patients, 15%) data are used as hold-out test purposes, this is our "internal" test-set.

We also evaluate QTNet on three "external" datasets; no fine-tuning was conducted. We test on the BWH dataset (3.1M ECG, 668k patients) and two Physionet datasets: RDVQ [12] and DMMLD [13]. The latter two datasets were generated for clinical trials exploring the effects of ion-channel blocking drugs on healthy subjects. The RDVQ study contains more than 5000 12-lead ECG from 22 subjects collected at a pre-administration and 15 post-administration time-markers for four drugs. Similarly, the DMMLD study collects more than 4000 12-lead ECG from 22 subjects over 14 time-markers for five drugs. Table 1 presents brief description of these datasets. The goal of this work is to evaluate the regression performance of QTNet on all those ECG data.

TABLE I. Dataset Properties

Dataset	MGH	BWH	RDVQ	DMMLD
Patients	903,593	667,060	22	22
ECG	4,223,689	3,171,283	5232	4211
Age (yr)	61 ± 18.6	60 ± 16.4	27 ± 5.4	26 ± 4.7
Female (%)	43.0	50.2	49.5	40.5
HR (bpm)	77 ± 20.3	77 ± 18.7	64 ± 9.5	67 ± 9.2
QT (ms)	394 ± 49.9	396 ± 47.6	400 ± 33.7	388 ± 24.3
QTc (ms)	439 ± 38.5	441 ± 35.3	412 ± 36.2	409 ± 27.4

B. Model Architecture and Training

QTNet is a single-channel residual neural network consisting of an ingest convolution layer, four residual blocks, each block of two convolution layers, and two fully-connected dense layers [Fig 1]. This architecture is adapted for regression tasks on single-lead ECG signals from existing pipelines [5]. 10-second ECG Lead-I signal with sampling rate of 250 Hz is input to the model as single-channel 1x2500 length tensor. The kernel size is 16 for all convolution layer filters. The ingest layer learns 64 filters with single sample stride. The four residual blocks learn convolution layers with 128, 196, 256, and 320 filters, consecutively. Each convolution layer is followed by batch normalization and ReLU activation, and the skip connections in each residual block are implemented with max pooling and 1-to-1 convolution layer. Average pooling is used to get 1x320 feature tensor from the output of the last residual block. The following two fully-connected layers learn the weights to regress two outputs: QTI and HR from this feature tensor. The pipeline is implemented with PyTorch.

QTNet is trained on the MGH training-validation dataset with the QT interval and HR labels. Each layer's weights are initialized from a normal distribution with variance depending on the layer size. The mean squared error (MSE) was used as the loss function that was minimized by the ADAM optimizer for regression. We use learning rate step scheduler to decay the rate in half every 3 epochs, starting from 0.01. With the batch size 512, each epoch consists of about 6000 iterations. The training is run for 100 epochs; we use "early-stopping" to reduce overfitting risk based on the validation loss. The computations are conducted on a workstation with 64-core processor, 512GB memory, and three 48GB GPUs.

C. ECG Delineation Algorithm

Given the lack of availability for open-source DL models for QTI regression from single-lead ECG, we use the popular NeuroKit2 [14] library to implement a QTI regression algorithm. Using built-in functions for peak-detection and fiducial point delineation, we calculate QTI and HR from the Lead-I ECG data for all the test-sets. This algorithm is used as the baseline method and the calculated QTI and HR are used for comparison against those inferred with QTNet.

D. Evaluation

QTNet is evaluated on four datasets: MGH test-set, BWH, RDVQ, and DMMLD datasets. We compare the inferred QTI to the expert-read labels using measurement agreement, correlation, and regression error metrics.

Correlation and Errors: Similarity between the inferred values and the labels is evaluated using the Pearson-R correlation coefficient. This coefficient ranges from -1 to 1; high and positive value refers to strong correlation. The regression

error is measured in mean absolute error (MAE), smaller error refers to better regression.

Bland-Altman Agreement: The Bland-Altman plot is a tool to quantify the amount of agreement between two measurement methods (the inferred QTI and the expert-read labels) for a variable. It provides two quantifiable metrics: the bias and the limit of agreement (LoA). The bias refers to the average difference between the two measurements, and the LoA represents the variation of those differences. The 95% LoA is defined as $1.96 \times SD_{diff}$, where SD_{diff} is the standard deviation of the difference in measurements.

From the inferred QTI and HR, we calculate QTc using the Bazett formula $QTc = QTI \times (HR / 60)^{1/2}$. This inferred QTc is compared against label using the aforementioned metrics.

III. Results

QTNet, without any fine-tuning, demonstrates robust regression performance across all four test-sets. Being trained on the MGH train-set, the performance of QTNet does not deteriorate when applied to "hold-out" or "unseen" datasets. As presented in Table 2, the MAE of QTNet remains notably lower than the baseline algorithm for all test-scenarios; moreover, the Pearson-R coefficient shows consistently strong correlation across datasets. While the baseline algorithm's performance significantly varies across datasets, QTNet performs consistently with MAE about 12.5 ms and Pearson-R 0.91, irrespective of the data sources. These results are also comparable to those reported in the literature on DL-based measurement of QT intervals from ECG signals [8-10].

The Bland-Altman analysis provides additional insight into the presence of systematic bias in the regression process. The 95% limit-of-agreement (LoA), which is 1.96 times the SD_{diff}, quantifies the reliability range for the predictions, especially in certain clinical applications. Smaller LoA refers to stronger agreement between the prediction and the true value. From Table 2, QTNet associated bias is small for the MGH test set and the BWH external dataset. Biases for the RDVQ and DMMLD are larger, but still significantly lower than that of the baseline model. Given the systematic differences in the data labeling procedures between the hospital systems and the Physionet studies, such difference in bias across sources is not unexpected, and is also observed for the baseline algorithm. The SD_{diff} ranges from 10.72 ms on the DMMLD data to 20.20 ms on the MGH test-set. These variances in difference are not only significantly lower than those from the baseline method, but also lower than those presented in related works [8-10]. For example, the best performance in SD_{diff} achieved by a DL model for QTc from single-lead ECG was 23 ms [10]. Similar result of 23.5 ms best SDdiff was reported by [8].

To investigate the spread of the measurement error, we use the Bland-Altman plots of the regression result on the RDVQ and DMMLD datasets, as shown in Fig 2. The plots show the distribution of the inference errors for QTNet and the baseline algorithm. QTNet shows positive biases for both datasets, as shown by the red-lines (labeled as 'mean' of the difference) on the plots. These biases indicate the tendency of the systematic error in QTNet and quantify possible adjustment to the predictions that can be made for better regression performance on such data. For both datasets, we notice the narrow spreads of the error distributions for QTNet, leading to small LoA

TABLE II. Performance Comparison

Data	Methods	Metrics			
		MAE (ms)	***Pearson-R (ratio)***	***Bias (ms)***	***SD_{diff} (ms)***
MGH Test-set N=624,652	***Baseline***	86.54	0.382	-82.37	84.28
	QTNet	12.63	0.914	0.59	20.20
BWH N=3,171,283	***Baseline***	90.79	0.368	-86.98	85.56
	QTNet	12.30	0.922	-0.33	18.61
RDVQ N=5,217	***Baseline***	38.14	0.345	-4.87	60.19
	QTNet	15.81	0.899	11.09	14.73
DMMLD N=4,211	***Baseline***	22.29	0.620	-8.36	30.20
	QTNet	9.39	0.899	4.60	10.72

result; the green-lines show the LoA on Fig 2. Such narrow LoA band highlights that the regression errors are bound within a small range, leading to increased reliability of the method. The LoA for the baseline model, on the other hand, are much wider, emphasizing the lower reliability of this method for many applications. These plots also underscore the consistency of QTNet in regression across datasets.

Similarly, the inferred QTc value from the QTNet inferred QTI and HR is compared against corresponding labels. We find similar performance for QTc as the QTI regression. On the MGH test-set, the MAE between QTNet inferred QTc and the label is 14 ms with a Pearson-R correlation coefficient of 0.8; on the BWH data, these metrics remain the same. On the RDVQ data, the MAE of 15.75 ms with Pearson-R 0.94 are reported for QTNet. And, for the DMMLD dataset, the QTc regression MAE is 9.73 ms and Pearson-R coefficient 0.93. These results highlight the improvement in performance of QTNet over those reported in literature [8-10].

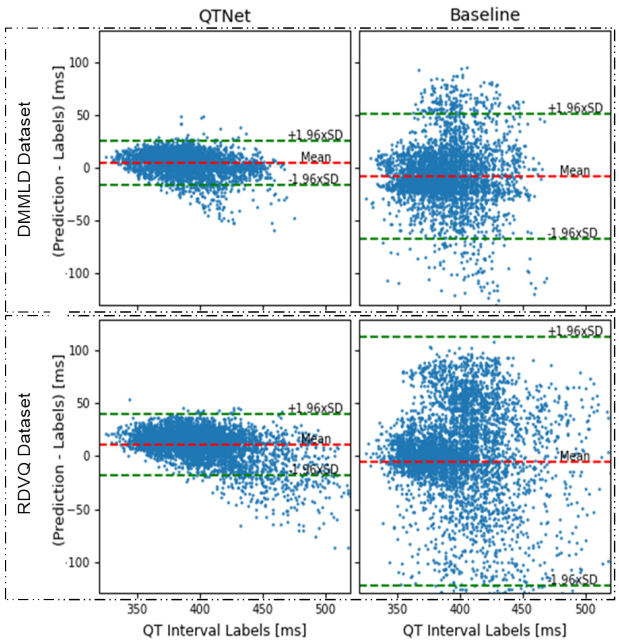

Figure 2. Bland-Altman agreement plots comparing QTNet and baseline method for measuring QT intervals on datasets DMMLD and RDVQ.

IV. Related Works

Recent DL models for ECG analysis have notably advanced in learning useful representations for identifying cardiac abnormalities and conditions, as well as in capturing the relationship between the ECG and other hemodynamic parameters. Representations learnt from large amount of ECG have been fine-tuned for multiple prediction tasks like atrial fibrillation, ventricular hypertrophy, heart blocks, and arrhythmias [5,6]. Decision support systems such as hypertension diagnosis are also implemented from such representations [7]. Though these models have been evaluated for classification tasks, they seem to show resilience against noise variance and intramodal variations. In addition, these studies lend credence to the application of ResNet as an attractive architecture for building DNNs for ECG-based classification or regression tasks. Hence, we design QTNet from 1d ResNet and adapt for regression of continuous variables.

ECG interval measurement using DL models from 12-lead or less is still an area of active research. Due to the challenges of regression learning, others have attempted to measure QTI not only as continuous variable regression but also as secondary to beat-length delineation or range estimation as multiclass classification [8-10]. For 12-lead ECG based DNN models, ResNet architectures have been employed to estimate the probability of an interval among small bins over the entire range; the performance shows SD_{diff} about 23 ms [8]. Whereas, [9] used a DNN to regress the intervals directly from 12-lead ECG, achieving the variance of difference reported in SD_{diff} of about 16 ms. For multilead inputs (Lead-I and II), this variance in regression error tends to be higher, as SD_{diff} is about 25 ms in [8] and 22 ms in [9]. Similar result has been reported for Cardiologs proprietary DL model on smartwatch ECG signals [10]. They also use a ResNet to delineate each beat and secondarily calculate QTI and QTc. Taking the guidance from these existing works, QTNet attempts to improve the performance and achieve generalizability.

Commercially available digital electrocardiograms use proprietary algorithms for automated QTI measurement from 12-lead ECG. The agreement, and the lack of it, among such algorithms varies significantly across cohorts. For QTI, pairwise mean differences between algorithms can range up to 10-13 ms [15]. Moving from 12-lead to lower number of leads reduces the accuracy of such algorithms even further; for example, AliveCor Lead-I measurements has shown SD_{diff} of 46 ms against GE measurements [3]. These challenges are observed not only across algorithms, the agreement among expert-reads seems to suffer significantly when reading the QTc from Apple watch Lead-I compared to those from 12-lead GE ECG; SD_{diff} 30 ms, median absolute error 18 ms [4]. These high differences show "allowable" limits toward clinical significance, while also motivate the need for complex data-driven models as objective high-throughput measurement tools.

V. Conclusion

This work presents a novel regression model for measuring QT interval from Lead-I ECG signal, and demonstrates strong agreement of the predicted values with corresponding 12-lead labels on four hold-out datasets. The model outperforms the baseline algorithm in all metrics overwhelmingly. The results also show significant improvement from those published in recent works. Consistently 'good' inference on all test-sets, without any fine-tuning, emphasizes the robustness and generalizability of this model to real-world datasets.

The metrics of evaluation also highlight the potential of applying QTNet for applications in remote QTI monitoring. Prolongation of QT secondary to drug effects are quantified with an increase in QTI by about 60 ms [8,12,13], which is much higher than the 95% limit-of-agreement of 20 ms for QTNet inferred values, and suggests that the method may be useful for real time tracking of QT prolongation. Two of the external datasets that we evaluated QTNet on, RDVQ and DMMLD, were acquired as part of studies focusing on drug-induced ECG changes with time after administration of various ion-channel blocking drugs. Future work will use those temporal data to evaluate the applicability of QTNet for identifying and predicting QT prolongation incidents for preventive purposes. Additionally, prospective studies of the method, in patients being given antiarrhythmic drugs that can lead to QT prolongation, are needed to ensure that the method can truly identify patients at high risk of significant QT prolongation and related drug-induced arrythmias.

References

[1] J. Tisdale, et al., "Development and validation of a risk score to predict QT interval prolongation in hospitalized patients," Circulation: Cardiovascular Quality and Outcomes, 2013, 6(4), pp. 479-87.

[2] J. Giudicessi, et al., "The QT interval: An emerging vital sign for the precision medicine era?" Circulation, 2019, 139(24), pp. 2711-3.

[3] P. Garabelli, et al., "Comparison of QT interval readings in normal sinus rhythm between a smartphone heart monitor and a 12-lead ECG for healthy volunteers and inpatients receiving Sotalol or Dofetilide," J. of Cardiovascular Electrophysiology, 2016, 27(7), pp. 827-32.

[4] M. Strik, et al., "Validating QT-interval measurement using the Apple Watch ECG to enable remote monitoring during the COVID-19 pandemic," Circulation, 2020,142(4), pp. 416-8.

[5] N. Diamant, et al., "Patient contrastive learning: A performant, expressive, and practical approach to electrocardiogram modeling," PLoS Comp. Biology, 2022, 18(2), e1009862.

[6] A. Ribeiro, et al., "Automatic diagnosis of the 12-lead ECG using a deep neural network," Nature Comm, 2020,11(1), pp. 1760.

[7] D. Schlesinger, et al., "A deep learning model for inferring elevated pulmonary capillary wedge pressures from the 12-lead electrocardiogram," JACC: Advances, 2022, 1(1), pp. 100003.

[8] J. Giudicessi, et al., "Artificial intelligence–enabled assessment of the heart rate corrected QT interval using a mobile electrocardiogram device," Circulation, 2021, 143(13), pp. 1274-86.

[9] M. Schram, et al., "Prediction of the heart rate corrected QT interval (QTc) from a novel, multilead smartphone-enabled ECG using a deep neural network," J. Amer. College of Cardiology, 2019, pp. 368.

[10] B. Maille, et al., "Smartwatch electrocardiogram and artificial intelligence for assessing cardiac-rhythm safety of drug therapy in COVID-19 pandemic. The QT-logs study," Int. J. of Cardiology, 2021, 331 pp. 333.

[11] A. Goldberger, et al. "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals," Circulation [Online], 2000, 101 (23), pp. e215–e220.

[12] L. Johannesen, et al., "Differentiating drug-induced multichannel block on the electrocardiogram: Randomized study of Dofetilide, Quinidine, Ranolazine, and Verapamil," Clin. Pharma. & Therap., 2014, pp. 549.

[13] L. Johannesen, et al., "Late Sodium current block for drug-induced long QT syndrome: Results from a prospective clinical trial," Clin. Pharma. & Therapeutics, 2016, 99(2), pp. 214-23.

[14] D. Makowski, et al., "NeuroKit2: A Python toolbox for neurophysiological signal processing," Behavior Research Methods, 2021, pp. 1689.

[15] P. Kligfield, et al., "Comparison of automated interval measurements by widely used algorithms in digital electrocardiographs," Amer. Heart Journal, 2018, 200, pp. 1.