Active Stacking for Heart Rate Estimation

Dongrui Wu, Chenfeng Guo
School of Artificial Intelligence and Automation
Huazhong University of Science and Technology
Wuhan, China
Email: {drwu, cfguo}@hust.edu.cn

Feifei Liu, Chengyu Liu
School of Instrument Science and Engineering
Southeast University
Nanjing, China
Email: feifeiliu1987@gmail.com, chengyu@seu.edu.cn

***Abstract*—Heart rate estimation from electrocardiogram signals is very important for the early detection of cardiovascular diseases. However, due to large individual differences and varying electrocardiogram signal quality, there does not exist a single reliable estimation algorithm that works well on all subjects. Every algorithm may break down on certain subjects, resulting in a significant estimation error. Ensemble regression, which aggregates the outputs of multiple base estimators for more reliable and stable estimates, can be used to remedy this problem. Moreover, active learning can be used to optimally select a few trials from a new subject to label, based on which a stacking ensemble regression model can be trained to aggregate the base estimators. This paper proposes four active stacking approaches, and demonstrates that they all significantly outperform three common unsupervised ensemble regression approaches, and a supervised stacking approach which randomly selects some trials to label. Remarkably, our active stacking approaches only need three or four labeled trials from each subject to achieve an average root mean squared estimation error below three beats per minute, making them very convenient for real-world applications. To our knowledge, this is the first research on active stacking, and its application to heart rate estimation.**

***Index Terms*—Active learning, ensemble regression, stacking, heart rate estimation**

I. Introduction

Cardiovascular diseases are the leading cause of human death. According to the World Health Organization [1], cardiovascular diseases take 17.9 million lives every year, accounting for 31% of all global deaths.

Electrocardiogram (ECG) is very useful in early detection of cardiovascular diseases. Recent years have witnessed rapid developments of wearable ECG systems for continuous ECG monitoring [2]. In such systems, real-time accurate heart rate estimation is critical to cardiovascular disease detection and treatment [3].

Unfortunately, ECG from wearable systems generally has poor quality due to bad electrode contact, wrong electrode positioning, body movements, and various noise [4]. As a result, traditional heart rate estimation algorithms, which mainly considered clinic quality ECG signals, may have difficulty on these wearable ECG systems. For example, Liu *et al.* [5] systematically evaluated ten widely used QRS detection algorithms in different application scenarios in six internationally recognized databases. Results showed that for the clinical ECG, whether it was normal or arrhythmic, the F1 measure of all algorithms was higher than 95%. However, the highest F1 score for wearable ECG was only 80.43%. A possible remedy is to perform ECG signal quality assessment [3] before further analysis, i.e., divide the ECG signal into acceptable and unacceptable parts, and discard the unacceptable portion. However, the discarded part may also contain valuable cardiovascular disease information.

Additionally, even when the ECG signal quality is satisfactory, due to large individual differences, there may not exist a single heart rate estimation algorithm that works well on all subjects. This paper considers how to use advanced machine learning approaches to cope with these problems.

Ensemble regression [6] has been frequently used to improve the estimation performance, by integrating multiple base estimators. In heart rate estimation, different QRS detectors can be viewed as base estimators. They are developed based on different ECG characteristics (e.g., power, amplitude, slope, curve length, etc [7]) and different detection methods (e.g., filtering, threshold setting, feature extraction, and post-processing [8]), and hence satisfy the basic requirement on the base learners in ensemble learning: diversity.

More specifically, we assume M base estimators are used to estimate the heart rates of N ECG trials from a particular subject. According to whether labeled training data are available or not, there are two types of ensemble regression approaches:

1) *Unsupervised ensemble regression*, where no labeled ECG trials are available. The simplest, maybe also the most frequently used, unsupervised ensemble regression approaches are to take the average or median of the M base estimators. However, as it will be shown later in this paper, because of individual differences, these approaches do not work well on heart rate estimation.
2) *Supervised ensemble regression*, where some labeled ECG trials are available. Some sophisticated supervised ensemble regression approaches [9], e.g., bagging [10], boosting [11], random forests [12], etc, require a relatively large number of labeled data. The simplest supervised ensemble regression approach, which also does not require too many labeled data, may be stacking [13], i.e., the final estimator is a weighted average of the base estimators, where the weights are computed from the labeled training data. Again, as it will be shown later in this paper, because of individual differences, it is very

This research was supported by the Hubei Technology Innovation Platform Grant 2019AEA171 and the National Natural Science Foundation of China Grant 61873321.

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

challenging to find a set of weights that fits all subjects. Usually some subject-specific labeled ECG trials must be obtained, based on which a subject-specific ensemble regression approach can then be designed to achieve a high estimation accuracy.

Intuitively, supervised ensemble regression would outperform unsupervised ensemble regression, if high-quality subject-specific labeled ECG trials can be acquired. Generally, the more such trials there are, the higher the estimation accuracy will be. However, for practical considerations, we'd like to minimize the number of subject-specific labeled ECG trials, as labeling each such trial requires an expert to visually examine and count the number of QRS waves in the ECG trial, which is both tedious and labor-intensive. So, it is desirable to reduce the number of subject-specific labeled ECG trials.

Active learning [14] is a popular and effective approach for this purpose. It deliberately selects a small number of most beneficial trials from the N unlabeled trials to label, so that a model trained from these labeled trials can achieve the best possible performance. Our previous research has demonstrated the outstanding performance of active learning in both classification [15], [16] and regression [17]–[19] tasks, in a variety of application domains. However, to our knowledge, no one has integrated stacking and active learning for heart rate estimation.

This paper proposes four active stacking approaches for estimator aggregation, which integrate active learning for regression (ALR) [17]–[19] and stacking. The idea is to use ALR to select a small number of most beneficial unlabeled trials, query an expert for their outputs, and then train a stacking model on them. We demonstrate their outstanding performances on heart rate estimation from ECG signals on 95 subjects: our active stacking approaches only need three or four labeled ECG trials from each subject to achieve an average root mean squared error below three beats per minute (bpm), making them very practical for real-world applications.

The remainder of this paper is organized as follows: Section II proposes four ALR approaches. Section III proposes four active stacking approaches. Section IV compares their performances on heart rate estimation from ECG signals. Finally, Section V draws conclusion.

II. Active Learning for Regression (ALR)

This section introduces four ALR approaches. The first two are unsupervised, whereas the last two are supervised.

Assume a subject has N ECG trials, each with its heart rate estimates $\mathbf{x}_n = [x_{n,1}, ..., x_{n,M}]$ from M base estimators, but initially none of these trials has a reference heart rate label. The goal of ALR is to optimally select K trials to label, so that an accurate regression model can be constructed from them to estimate the heart rate for the remaining $N - K$ trials.

A. GSx

Yu and Kim [20] proposed a greedy sampling (GS) ALR approach, which selects the trials to label based entirely on their locations in the input space. Thus, it does not need any label information at all. However, the original GS approach did not explain how the first trial was selected. We [19] recently introduced GSx to accommodate this. GSx is essentially the same as GS, except that it also includes a strategy to select the first trial for labeling.

GSx selects the first trial as the one whose $\mathbf{x}_n$ is the closest to the centroid of all N $\mathbf{x}_n$ (i.e., the one with the shortest mean distance to the remaining $N-1$ $\mathbf{x}_n$), and the remaining $K-1$ trials incrementally. In this way, the first selected trial is the most representative one in the N trials.

Without loss of generality, assume the first k trials $\{\mathbf{x}_l\}_{l=1}^{k}$ have already been selected. For each of the remaining $N-k$ unlabeled trials $\{\mathbf{x}_n\}_{n=k+1}^{N}$, GSx computes first its distance to each of the k labeled trials:

$$d_{nl}^{\mathbf{x}} = ||\mathbf{x}_n - \mathbf{x}_l||, \quad l = 1, ..., k; \; n = k+1, ..., N \tag{1}$$

then $d_n^{\mathbf{x}}$, the shortest distance from $\mathbf{x}_n$ to all k labeled trials:

$$d_n^{\mathbf{x}} = \min_l d_{nl}^{\mathbf{x}}, \quad n = k+1, ..., N \tag{2}$$

and finally selects the trial with the maximum $d_n^{\mathbf{x}}$ to label.

In summary, GSx selects the first trial as the one closest to the centroid of the pool, and each subsequent trial farthest away from all previously selected ones in the input space, to achieve the maximum diversity among the selected trials.

B. RD

We [17] recently proposed a representativeness-diversity (RD) approach for ALR. As its name suggests, it considers both representativeness and diversity in all trial selections. In contrast, GSx considers only representativeness in selecting the first trial, and only diversity in subsequent selections.

RD selects all K trials simultaneously. It performs k-means ($k = K$) clustering on the N unlabeled trials, and then selects from each cluster the trial closest to the cluster centroid for labeling. This selection strategy ensures representativeness, because each trial is a good representation of the cluster it belongs to. It also ensures diversity, because these K clusters cover the full input space of $\mathbf{x}_n$, and they do not overlap.

As GSx, RD does not need any reference label information at all, so it is a completely unsupervised ALR approach.

C. RD-EMCM

RD only considers representativeness and diversity. However, as pointed out in [17], informativeness is also an essential criterion in ALR. An RD-EMCM ALR approach was proposed in [17], which considers also the informativeness through expected model change maximization (EMCM) [21].

RD-EMCM first uses RD to select $K_0 = 2$ trials, and queries for their reference labels. To select the next trial to label, it performs k-means ($k = K_0 + 1$) clustering on the N trials, and identifies the largest cluster that does not already contain any labeled trial. It will then select the $(K_0 + 1)$th trial from this cluster. However, instead of selecting the one closest to its centroid, as in RD, now it uses EMCM to select the most informative trial to label. The details of EMCM are given next.

EMCM first uses all labeled trials to build a linear regression model (e.g., ridge regression, or linear SVR). Let its estimated heart rate for the nth trial be $\hat{y}_n$. EMCM then uses bootstrap to construct another P linear regression models from the labeled trials. Let the pth model's estimated heart rate for the nth trial be $\hat{y}_n^p$. Then, for each unlabeled trials, EMCM computes [21]

$$g(\mathbf{x}_n) = \frac{1}{P}\sum_{p=1}^{P} \|(\hat{y}_n^p - \hat{y}_n)\mathbf{x}_n\|, \tag{3}$$

and selects the trial with the maximum $g(\mathbf{x}_n)$ to label.

RD-EMCM is a supervised ALR approach, because it needs the reference labels to train the regression models in EMCM.

D. iGS

Improved greedy sampling (iGS) is a supervised ALR approach proposed in [19], which is supposed to improve GSx by considering also feature selection/weighting.

iGS first uses GSx to select the initial $K_0 = 2$ trials to label. Assume the first k trials $\{\mathbf{x}_l\}_{l=1}^k$ have already been labeled with true heart rates $\{y_l\}_{l=1}^k$. For each of the remaining $N-k$ unlabeled trials $\{\mathbf{x}_n\}_{n=k+1}^N$, iGS computes:

$$d_{nl}^{\mathbf{x}} = ||\mathbf{x}_n - \mathbf{x}_l|| \tag{4}$$

$$d_{nl}^{y} = |f(\mathbf{x}_n) - y_l| \tag{5}$$

$$d_n^{\mathbf{x}y} = \min_l d_{nl}^{\mathbf{x}} d_{nl}^{y} \tag{6}$$

and then selects the trial with the maximum $d_n^{\mathbf{x}y}$, i.e., the trial located farthest away from all previously selected trials in both input and output spaces, to label.

III. Active Stacking

Stacking requires some labeled trials, whereas ALR can optimally select a small number of trials to label. So, it's natural to integrate them for better performance. Four active stacking approaches are proposed next.

A. AS-GSx

AS-GSx integrates stacking and GSx. It uses GSx to select K trials to query for their reference heart rates, and then checks if any base estimator has the same heart rate estimates as the reference for all K selected trials. If yes, then for each of the remaining $N-K$ trials, the median of these base estimators is taken as its final estimate. Otherwise, it trains a linear SVR model from the K labeled trials as the final stacking model.

The pseudo-code of AS-GSx is given in Algorithm 1.

B. AS-RD

AS-RD integrates stacking and RD. It's almost identical to AS-GSx, except that GSx is replaced by RD as the ALR approach. Its pseudo-code is given in Algorithm 2.

Algorithm 1: The AS-GSx active stacking approach.

Input: N unlabeled trials, $\{\mathbf{x}_n\}_{n=1}^N$;
K, the maximum number of labels to query.
Output: The stacking regression model $f(\mathbf{x})$.
Set $Z = \{\mathbf{x}_n\}_{n=1}^N$, and $S = \emptyset$;
Identify $\mathbf{x}'$, the trial closest to the centroid of Z;
Move $\mathbf{x}'$ from Z to S;
Re-index the trial in S as $\mathbf{x}_1$, and the trials in Z as $\{\mathbf{x}_n\}_{n=2}^N$;
for $k = 1, ..., K-1$ **do**
  for $n = k+1, ..., N$ **do**
    Compute $d_n^{\mathbf{x}}$ in (2);
  end
  Identify the $\mathbf{x}'$ that has the largest $d_n^{\mathbf{x}}$;
  Move $\mathbf{x}'$ from Z to S;
  Re-index the trials in S as $\{\mathbf{x}_l\}_{l=1}^{k+1}$, and the trials in Z as $\{\mathbf{x}_n\}_{n=k+2}^N$;
end
Query to label all K trials in S;
if *There exist some base estimators which give identical estimates to the true labels in* S **then**
  $f(\mathbf{x})$ is the median of these base estimator outputs;
else
  Construct a linear SVR model $f(\mathbf{x})$ from S.
end

Algorithm 2: The AS-RD active stacking approach.

Input: N unlabeled trials, $\{\mathbf{x}_n\}_{n=1}^N$;
K, the maximum number of labels to query.
Output: The stacking regression model $f(\mathbf{x})$.
Perform k-means clustering on $\{\mathbf{x}_n\}_{n=1}^N$, where $k = K$;
Select from each cluster the trial closest to its centroid, and query for its label;
if *There exist some base estimators which give identical estimates to the true labels for all* K *trials* **then**
  $f(\mathbf{x})$ is the median of these base estimator outputs;
else
  Construct a linear SVR model $f(\mathbf{x})$ from the K labeled trials.
end

C. AS-RD-EMCM

AS-RD-EMCM integrates stacking and RD-EMCM. It first uses RD-EMCM to select $K_0 = 2$ trials to query for their reference heart rates, and trains a linear SVR stacking model from them. This model can then be used in RD-EMCM to select the next trial to label, and the linear SVR stacking model is then updated. This process iterates until K trials have been selected and labeled. Finally, we check if any base estimator has the same heart rate estimates as the reference for all K selected trials. If yes, then for each of the remaining $N-K$

trials, the median of these base estimators is taken as the final estimate. Otherwise, we train a linear SVR model from the K labeled trials as the final stacking model.

The pseudo-code of AS-RD-EMCM is given in Algorithm 3.

Algorithm 3: The AS-RD-EMCM active stacking approach.

Input: N unlabeled trials, $\{\mathbf{x}_n\}_{n=1}^N$;
K, the maximum number of labels to query.
Output: The stacking regression model $f(\mathbf{x})$.
Perform k-means clustering on $\{\mathbf{x}_n\}_{n=1}^N$, where $k = 2$;
Select from each cluster the trial closest to its centroid, and query for its label;
Construct a linear SVR model $f(\mathbf{x})$ from the two labeled trials;
for $k = 3, ..., K$ **do**
  Perform k-means clustering on $\{\mathbf{x}_n\}_{n=1}^N$;
  Identify the largest cluster that does not already contain any labeled trial;
  Compute $g(\mathbf{x}_n)$ in (3) for each trial in the above cluster;
  Select the trial with the maximum $g(\mathbf{x}_n)$ to label;
  Construct a linear SVR model $f(\mathbf{x})$ from the k labeled trials;
end
if *There exist some base estimators which give identical estimates to the true labels for all K trials* **then**
  $f(\mathbf{x})$ is the median of these base estimator outputs;
else
  Construct a linear SVR model $f(\mathbf{x})$ from the K labeled trials.
end

D. AS-iGS

AS-iGS integrates stacking and iGS. It's almost identical to AS-RD-EMCM, except that RD-EMCM is replaced by iGS as the ALR approach. Its pseudo-code is given in Algorithm 4.

IV. Experiments and Results

A. Datasets

One hundred ECG recordings in the augmented training set of the 2014 PhysioNet/CinC Challenge [22], available freely on the PhysioNet platform, were used in this study. They were from patients with a wide range of problems as well as healthy volunteers. Each recording was 10 minutes or shorter, sampled at 360 Hz with 16-bit resolution. Four recordings (2041, 2728, 41024, 41778) shorter than 2 minutes, and one consisting of pure Gaussian noise (42878), were excluded. The remaining 95 ECG recordings had manually annotated QRS complex locations. Heart rates calculated from these locations were used as the references for algorithm evaluations. Most subjects had close to 120 trials, but a few had less than 40. On average each subject had 108.5 trials.

Algorithm 4: The AS-iGS active stacking approach.

Input: N unlabeled trials, $\{\mathbf{x}_n\}_{n=1}^N$;
K, the maximum number of labels to query.
Output: The stacking regression model $f(\mathbf{x})$.
Set $Z = \{\mathbf{x}_n\}_{n=1}^N$, and $S = \emptyset$;
Identify $\mathbf{x}'$, the trial closest to the centroid of Z;
Move $\mathbf{x}'$ from Z to S;
Re-index the trial in S as $\mathbf{x}_1$, and the trials in Z as $\{\mathbf{x}_n\}_{n=2}^N$;
for $n = 2, ..., N$ **do**
  Compute $d_n^{\mathbf{x}}$ in (2);
end
Identify the $\mathbf{x}'$ that has the largest $d_n^{\mathbf{x}}$;
Move $\mathbf{x}'$ from Z to S;
Re-index the trials in S as $\{\mathbf{x}_l\}_{l=1}^2$, and the trials in Z as $\{\mathbf{x}_n\}_{n=3}^N$;
Query to label the two trials in S;
Construct a linear SVR model $f(\mathbf{x})$ from S;
for $k = 3, ..., K$ **do**
  for $n = k, ..., N$ **do**
    Compute $d_n^{\mathbf{x}y}$ in (6);
  end
  Identify the $\mathbf{x}'$ that has the largest $d_n^{\mathbf{x}y}$;
  Move $\mathbf{x}'$ from Z to S;
  Query to label $\mathbf{x}'$ in S;
  Re-index the trials in S as $\{\mathbf{x}_l\}_{l=1}^k$, and the trials in Z as $\{\mathbf{x}_n\}_{n=k+1}^N$;
  Update the linear SVR model $f(\mathbf{x})$ using S.
end
if *There exist some base estimators which give identical estimates to the true labels in S* **then**
  $f(\mathbf{x})$ is the median of these base estimator outputs;
else
  Construct a linear SVR model $f(\mathbf{x})$ from S.
end

B. Base Estimators

The following 12 QRS detection algorithms were used as our base estimators [5]:

1) Pan-Tompkins [23], which has been widely used as a baseline QRS detection algorithm.
2) Hamilton-Tompkins-mean [24], which is an improvement to the Pan-Tompkins algorithm.
3) Hamilton-Tompkins-median [24], which is another improvement to the Pan-Tompkins algorithm.
4) RS-slope [25], which uses the RS slope to detect the QRS complexes.
5) Sixth-power [26], which relies on the sixth power of the ECG signal to identify the QRS complexes.
6) Finite state machine (FSM) [27], which uses a dynamic finite state machine based threshold to detect the R peaks.
7) Improved FSM (iFSM) [5], which improves parameter selection and threshold estimation in FSM.

8) U3 [28], which uses the U3 transform (a non-linear time-domain transform) for QRS detection.
9) Difference operation algorithm (DOM) [29], which uses the positive and negative extremes of the low-pass filtered differential ECG signal to detect the R peaks.
10) jqrs [30], which fuses R peaks detected from the ECG using an energy detector with those from the arterial blood pressure waveform using the length transform.
11) Optimized knowledge-based method (OKM) [31], which detects QRS complexes in ECG signals based on two event-related moving-average filters.
12) UNSW [3], which generates a feature signal containing information of ECG amplitude and derivative, and then performs filtering and adaptive thresholding.

Boxplots of the root mean squared errors (RMSEs) of the 12 base estimators on the 95 subjects are shown in Fig. 1. Due to large individual differences, every base estimator broke down on certain subjects, giving heart rate estimates zero or over 1000 bpm, and hence very large RMSEs. The mean and standard deviation of the RMSEs of the 12 base estimators are shown in the first part of Table I. Among the 12 base estimators, sixth-power achieved the smallest average RMSE (10.55 bpm), and RS-slope the largest (29.57 bpm). Given that the average reference heart rate across the 95 subjects was 87.99 bpm, these RMSEs represented 11.99-33.61% relative error. According to ANSI/AAMI[1] EC13:2002, which establishes minimum safety and performance requirements for cardiac monitors, heart rate meters, and alarms: "*the minimum allowable heart rate meter range shall be 30 bpm to 200 bpm, with an allowable readout error of no greater than 10 percent of the input rate or 5 bpm, whichever is greater.*" Clearly, all 12 base estimators have errors exceeding the 10 percent or 5 bpm threshold, and hence may not be suitable for real-world applications.

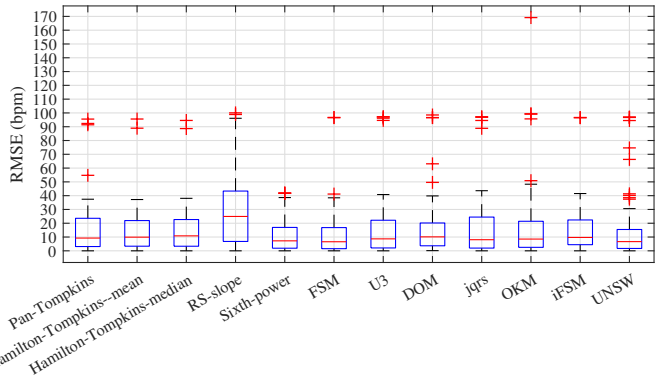

Fig. 1. Boxplots of the RMSEs of the 12 base estimators on the 95 subjects.

In summary, we have shown that the base estimators were very unstable, and none of them may be used for heart rate estimation alone in practice.

[1]https://webstore.ansi.org/standards/aami/ansiaamiec132002.

TABLE I
THE MEAN AND STANDARD DEVIATION (STD) OF THE RMSES OF DIFFERENT APPROACHES.

Category		Approach	RMSE mean	RMSE std
Base Estimator		Pan-Tompkins	15.49	18.06
		Hamilton-Tompkins-mean	14.69	15.78
		Hamilton-Tompkins-median	14.87	15.73
		RS-slope	29.57	26.92
		Sixth-power	**10.55**	**10.95**
		FSM	12.14	16.16
		iFSM	15.26	16.54
		U3	15.68	20.47
		DOM	15.67	19.03
		jqrs	16.33	20.83
		OKM	17.09	25.30
		UNSW	14.22	21.88
Unsupervised Ensemble Regression		LOSO-CV	**11.37**	**11.65**
		Average	11.97	12.14
		Median	12.10	16.86
Supervised Stacking	$K=2$	RS	5.55	4.45
		AS-GSx	**3.18**	**3.07**
		AS-RD	3.76	4.02
		AS-RD-EMCM	3.76	4.02
		AS-iGS	**3.18**	**3.07**
	$K=3$	RS	4.96	4.15
		AS-GSx	**2.97**	2.68
		AS-RD	2.98	**2.65**
		AS-RD-EMCM	3.12	2.67
		AS-iGS	2.99	2.66
	$K=4$	RS	4.64	3.97
		AS-GSx	**2.81**	**2.45**
		AS-RD	2.98	2.95
		AS-RD-EMCM	3.02	2.75
		AS-iGS	2.92	2.78
	$K=5$	RS	4.48	3.89
		AS-GSx	2.76	2.70
		AS-RD	**2.64**	**2.48**
		AS-RD-EMCM	2.90	2.58
		AS-iGS	2.99	3.05

C. Performances of the Unsupervised Ensemble Regression Approaches

Before testing our proposed active stacking algorithms, we would like to check first if unsupervised ensemble regression can work well. If so, then one should prefer unsupervised ensemble regression, since it does not require manually labeling some ECG trials for each new subject, and hence is very convenient to use.

In addition to *average* and *median*, *leave-one-subject-out cross-validation (LOSO-CV)* was also considered. From the 95 subjects, each time we selected one as the test subject, and the remaining 94 as training subjects. We combined trials from all 94 training subjects to train a stacking model (we tried both ridge regression with $\lambda = 0.01$ and linear SVR; however, the latter was too slow to converge, so we only report the ridge regression results), and computed its RMSE on the test subject. This process was repeated 95 times so that each subject acted as the test subject once. Note that this approach is *unsupervised for the new subject*, because we do not need any reference heart rates from him/her; however, it assumes that we know the reference heart rates of other subjects, so that the stacking model can be built. From this perspective, it is supervised on the existing subjects.

The mean and standard deviation of the RMSEs of the three algorithms on the 95 subjects are shown in the second part of Table I. Given that the mean heart rate across the 95 subjects was 87.99 bpm, these RMSEs represented $12.92 - 13.75\%$ relative error, which should not be acceptable in practice.

Boxplots of the RMSEs of the three unsupervised ensemble regression approaches on the 95 subjects are shown in Fig. 2. They were better than most base estimators, but still worse than the best base estimator.

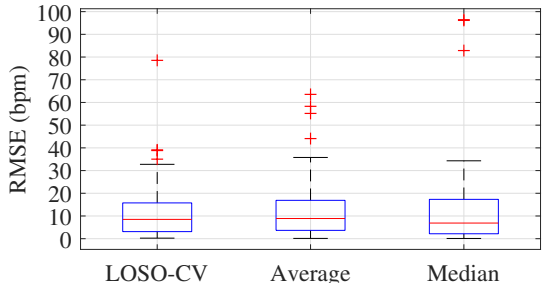

Fig. 2. Boxplots of the RMSEs of the three unsupervised ensemble regression approaches on the 95 subjects.

In summary, we have shown that, due to large individual differences, unsupervised ensemble regression approaches may not be accurate enough to be used for practical heart rate estimation.

D. Performances of the Supervised Stacking Approaches

Next we compare the performances of five supervised stacking algorithms: Random sampling (RS), AS-GSx, AS-RD, AS-RD-EMCM, and AS-iGS. The latter four have been introduced in Algorithms 1-4 in Section III. RS is similar to AS-GSx, except that GSx is replaced by random sampling.

Boxplots of the RMSEs of the five supervised stacking approaches on the 95 subjects are shown in Fig. 3, for different K. Clearly, these RMSEs were much smaller than those of the 12 base estimators (Fig. 1), and also much smaller than those of the three unsupervised ensemble regression approaches (Fig. 2).

Fig. 3 also shows that generally the RMSEs of all five supervised stacking approaches decreased with the increase of K. To visualize this more clearly, we plot the mean RMSEs of the five supervised stacking approaches across the 95 subjects in the left panel of Fig. 4, and also show them in the third part of Table I. Generally there was a decreasing trend for each approach, which is intuitive: the more labeled trials we have, the better a stacking model can be trained. Remarkably, the RMSEs of the four proposed active stacking approaches converged at $K = 3$ or $K = 4$, i.e., only three or four labeled trials were needed for these active stacking approaches to achieve a low RMSE, which is very favorable in practice.

The left subfigure of Fig. 4 also shows that the RMSEs of the four active stacking approaches were much smaller than those of RS. The right subfigure of Fig. 4 shows their ratios to the mean RMSE of RS. Compared with RS, each active stacking approach can reduce the RMSE by $35 - 40\%$, suggesting

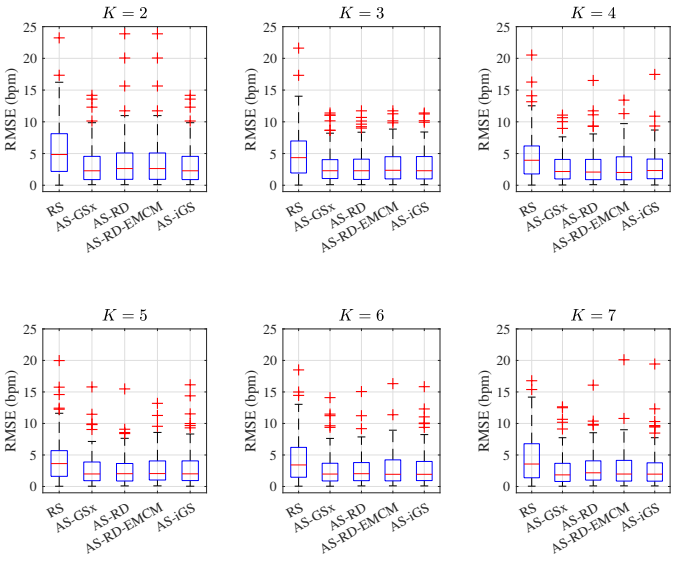

Fig. 3. Boxplots of the RMSEs of the five supervised stacking approaches, for different K.

the effectiveness of using ALR in heart rate estimation. The four active stacking approaches had similar performances.

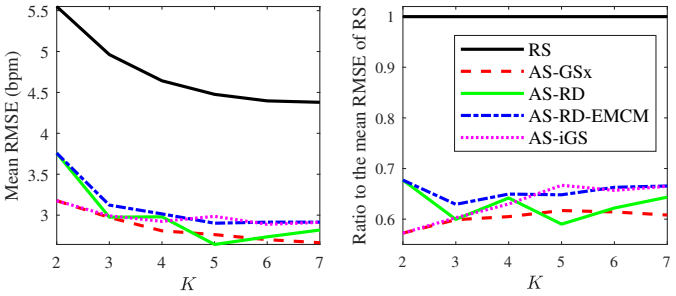

Fig. 4. Mean RMSEs (left) of the five supervised stacking approaches across the 95 subjects, and the ratio (right) to the mean RMSE of RS.

To find out if there were statistically significant differences between the five supervised stacking approaches, non-parametric multiple pairwise comparison tests using Dunn's procedure [32], with a p-value correction using the False Discovery Rate method [33], were performed on the 95×5 mean RMSEs (for each algorithm on each subject, we computed the mean RMSE for $K \in [2, 7]$). The results are shown in Table II, where the statistically significant ones are marked in bold. All four active stacking approaches significantly outperformed RS, but there were no statistically significant differences among the four active stacking approaches.

In summary, we have shown that all five supervised stacking approaches significantly outperformed the 12 base estimators, and the three unsupervised ensemble regression approaches. The four active stacking approaches further significantly outperformed supervised stacking by random sampling. So, active stacking is indeed effective in heart rate estimation.

TABLE II
p-VALUES OF NON-PARAMETRIC MULTIPLE COMPARISONS ON THE FIVE SUPERVISED STACKING APPROACHES ($p = 0.05$).

	RS	AS-GSx	AS-RD	AS-RD-EMCM
AS-GSx	**.0019**			
AS-RD	**.0034**	.5333		
AS-RD-EMCM	**.0043**	.5296	.4361	
AS-iGS	**.0019**	.4740	.4858	.4690

E. Discussions

In all four active stacking approaches (Algorithms 1-4), when there exist some base estimators whose outputs are identical to the reference heart rates on all selected trials, we take the median of these base estimators as the final output, instead of performing a linear SVR. This is because: 1) taking the median is intuitive, as the selected base estimators have identical performance on the reference trials, and hence they cannot be distinguished; 2) taking the median is much simpler than performing a linear SVR; and, 3) empirically taking the median[2] gave smaller RMSEs. Fig. 5 shows the average RMSEs of three variants of the algorithm:

1) *Median*, which takes the median of the selected base estimators.
2) *Subset*, which performs a linear SVR on the selected base estimators.
3) *All*, which performs a linear SVR on all 12 base estimators.

Taking the median had the smallest RMSEs for AS-GSx and AS-iGS, and comparable RMSEs with the two SVR approaches for AS-RD and AS-RD-EMCM (when $K \geq 3$). So, we used the median in our algorithms.

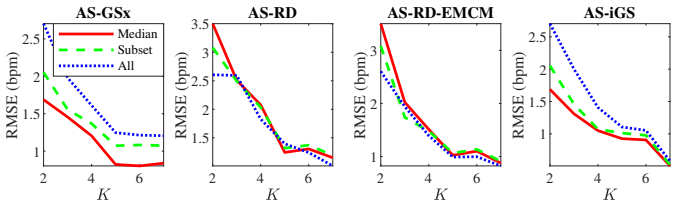

Fig. 5. Average RMSEs of three variants of the algorithm, when there exist some base estimators whose outputs are identical to the reference heart rates on all selected trials.

Intuitively, if there exist some base estimators whose outputs are identical to the reference heart rates on all selected trials, then these subjects may be easier to handle than others, i.e., they may have smaller RMSEs. To verify this, we show the RMSEs from these subjects (red dots, sorted in ascending order for easy visualization) versus those from the remaining subjects (black dots, sorted in ascending order for easy visualization) in Fig. 6. In each subfigure the vertical red (black) dashed line indicates the number of red (black) dots, and the horizontal red (black) dashed line indicates the mean RMSE of the red (black) dots. Each horizontal red line was always lower than the corresponding horizontal black line, confirming our hypothesis. As K increased, the number of red dots decreased (the corresponding vertical red line moved left), which is intuitive, because fewer base estimators were able to completely match the reference heart rates. However, as K increased, the horizontal red line also became lower (the RMSE was smaller), which is reasonable, as the survived subjects were easier to handle.

[2]We could also take the mean of the selected base estimators; however, it gave a larger RMSE than taking the median, because the mean is more sensitive to outliers than the median.

V. CONCLUSION

Heart rate estimation from ECG signals is very important for the early detection of cardiovascular diseases. However, due to large individual differences and varying ECG signal quality, there does not exist a single reliable estimation algorithm that works well on all subjects. Every algorithm may break down on certain subjects, resulting in a significant estimation error. Ensemble regression, which aggregates the outputs of multiple base estimators for more reliable and stable estimates, is a remedy to this problem. Additionally, active learning can be used to optimally select a few trials from a new subject to label, based on which a stacking ensemble regression model can be trained to properly aggregate the base estimators. This paper has proposed four active stacking approaches, and demonstrated that they all significantly outperformed three common unsupervised ensemble regression approaches, and a supervised stacking approach which randomly selects some trials to label. Remarkably, our active stacking approaches only need three or four labeled trials from each subject to achieve an average root mean squared estimation error below three bpm, making them very convenient for real-world applications. To our knowledge, this is the first research on active stacking, and its application to heart rate estimation.

REFERENCES

[1] "Cardiovascular disease," 2019, accessed Feb 16. [Online]. Available: https://www.who.int/cardiovascular_diseases/world-heart-day/en/

[2] C. Liu, X. Zhang, L. Zhao, F. Liu, X. Chen, Y. Yao, and J. Li, "Signal quality assessment and lightweight QRS detection for wearable ECG SmartVest system," *IEEE Internet of Things Journal*, 2019.

[3] H. Khamis, R. Weiss, Y. Xie, C. Chang, N. H. Lovell, and S. J. Redmond, "QRS detection algorithm for telehealth electrocardiogram recordings," *IEEE Trans. on Biomedical Engineering*, vol. 63, no. 7, pp. 1377–1388, 2016.

[4] Q. Li, R. G. Mark, and G. D. Clifford, "Robust heart rate estimation from multiple asynchronous noisy sources using signal quality indices and a Kalman filter," *Physiological Measurement*, vol. 29, no. 1, pp. 15–32, 2008.

[5] F. Liu, C. Liu, X. Jiang, Z. Zhang, Y. Zhang, J. Li, and S. Wei, "Performance analysis of ten common QRS detectors on different ECG application cases," *Journal of Healthcare Engineering*, vol. 2018, p. 9050812, 2018.

[6] Z.-H. Zhou, *Ensemble Methods: Foundations and Algorithms*. Boca Raton, FL: CRC press, 2012.

[7] M. Elgendi, B. Eskofier, S. Dokos, and D. Abbott, "Revisiting QRS detection methodologies for portable, wearable, battery-operated, and wireless ECG systems," *PLoS One*, vol. 9, no. 1, pp. e840 181–18, 2014.

[8] B. U. Kohler, C. Hennig, and R. Orglmeister, "The principles of software QRS detection," *IEEE Engineering in Medicine & Biology Magazine*, vol. 21, no. 1, pp. 42–57, 2002.

[9] P. Buhlmann, *Bagging, Boosting and Ensemble Methods*. Springer, 2010.

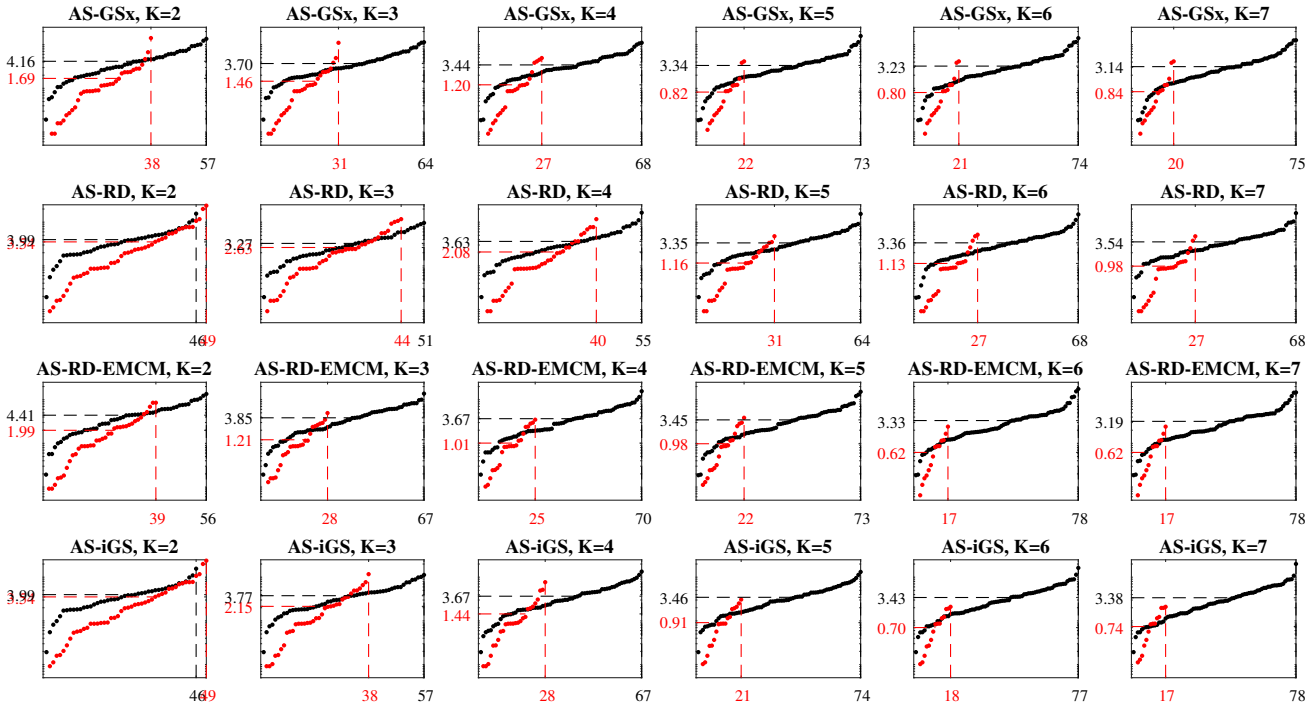

Fig. 6. Red dots: RMSEs of subjects who have some base estimators whose outputs are identical to the reference heart rates on all K selected trials. Black dots: RMSEs of the remaining subjects. Each dot represents one subject. Dots of the same color are sorted in ascending order for easy visualization. Vertical axis: RMSE in bpm (logarithmic scale is used to better distinguish between the values); horizontal axis: subject. In each subfigure the vertical red (black) dashed line indicates the number of red (black) dots, and the horizontal red (black) dashed line indicates the mean RMSE of the red (black) dots.

[10] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, pp. 123–140, 1996.

[11] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119–139, 1997.

[12] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 5, pp. 5–32, 2001.

[13] L. Breiman, "Stacked regressions," *Machine Learning*, vol. 24, pp. 49–64, 1996.

[14] B. Settles, "Active learning literature survey," University of Wisconsin–Madison, Computer Sciences Technical Report 1648, 2009.

[15] A. Marathe, V. Lawhern, D. Wu, D. Slayback, and B. Lance, "Improved neural signal classification in a rapid serial visual presentation task using active learning," *IEEE Trans. on Neural Systems and Rehabilitation Engineering*, vol. 24, no. 3, pp. 333–343, 2016.

[16] D. Wu, V. J. Lawhern, W. D. Hairston, and B. J. Lance, "Switching EEG headsets made easy: Reducing offline calibration effort using active wighted adaptation regularization," *IEEE Trans. on Neural Systems and Rehabilitation Engineering*, vol. 24, no. 11, pp. 1125–1137, 2016.

[17] D. Wu, "Pool-based sequential active learning for regression," *IEEE Trans. on Neural Networks and Learning Systems*, vol. 30, no. 5, pp. 1348–1359, 2019.

[18] D. Wu and J. Huang, "Affect estimation in 3D space using multi-task active learning for regression," *IEEE Trans. on Affective Computing*, 2020, in press.

[19] D. Wu, C.-T. Lin, and J. Huang, "Active learning for regression using greedy sampling," *Information Sciences*, vol. 474, pp. 90–105, 2019.

[20] H. Yu and S. Kim, "Passive sampling for regression," in *IEEE Int'l. Conf. on Data Mining*, Sydney, Australia, December 2010, pp. 1151–1156.

[21] W. Cai, Y. Zhang, and J. Zhou, "Maximizing expected model change for active learning in regression," in *Proc. IEEE 13th Int'l. Conf. on Data Mining*, Dallas, TX, December 2013.

[22] G. Moody, B. Moody, and I. Silva, "Robust detection of heart beats in multimodal data: the PhysioNet/Computing in Cardiology Challenge 2014," in *Proc. Computing in Cardiology Conference*. Cambridge, MA: IEEE, Sep. 2014, pp. 549–552.

[23] J. Pan and W. J. Tompkins, "A real-time QRS detection algorithm," *IEEE Trans. on Biomedical Engineering*, vol. 32, no. 3, pp. 230–236, 1985.

[24] P. S. Hamilton and W. J. Tompkins, "Quantitative investigation of QRS detection rules using the MIT/BIH arrhythmia database," *IEEE Trans. on Biomedical Engineering*, no. 12, pp. 1157–1165, 1986.

[25] P. Podziemski and J. Gieraltowski, "Fetal heart rate discovery: algorithm for detection of fetal heart rate from noisy, noninvasive fetal ECG recordings," in *Proc. Computing in Cardiology*, Zaragora, Sep. 2013, pp. 333–336.

[26] A. K. Dohare, V. Kumar, and R. Kumar, "An efficient new method for the detection of QRS in electrocardiogram," *Computers & Electrical Engineering*, vol. 40, no. 5, pp. 1717–1730, 2014.

[27] R. Gutiérrez-Rivas, J. J. García, W. P. Marnane, and A. Hernández, "Novel real-time low-complexity QRS complex detector based on adaptive thresholding," *IEEE Sensors Journal*, vol. 15, no. 10, pp. 6036–6043, 2015.

[28] M. Paoletti and C. Marchesi, "Discovering dangerous patterns in long-term ambulatory ECG recordings using a fast QRS detection algorithm and explorative data analysis," *Computer Methods and Programs in Biomedicine*, vol. 82, no. 1, pp. 20–30, 2006.

[29] T. De Cooman, G. Goovaerts, C. Varon, D. Widjaja, T. Willemen, and S. Van Huffel, "Heart beat detection in multimodal data using automatic relevant signal detection," *Physiological Measurement*, vol. 36, no. 8, pp. 1691–1704, 2015.

[30] A. E. Johnson, J. Behar, F. Andreotti, G. D. Clifford, and J. Oster, "Multimodal heart beat detection using signal quality indices," *Physiological Measurement*, vol. 36, no. 8, pp. 1665–1677, 2015.

[31] M. Elgendi, "Fast QRS detection with an optimized knowledge-based method: Evaluation on 11 standard ECG databases," *PloS One*, vol. 8, no. 9, p. e73557, 2013.

[32] O. J. Dunn, "Multiple comparisons using rank sums," *Technometrics*, vol. 6, pp. 214–252, 1964.

[33] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society, Series B (Methodological)*, vol. 57, pp. 289–300, 1995.