HypUC: Hyperfine Uncertainty Calibration with Gradient-boosted Corrections for Reliable Regression on Imbalanced Medical Time Series

Anonymous authors
Paper under double-blind review

Abstract

The automated analysis of medical time series, such as the electrocardiogram (ECG), electroencephalogram (EEG), pulse oximetry, etc, has the potential to serve as a valuable tool for diagnostic decisions, allowing for remote monitoring of patients and more efficient use of expensive and time-consuming medical procedures. Deep neural networks (DNNs) have been demonstrated to be effective in processing such signals. However, previous research has largely focused on classifying medical time series rather than attempting to regress the continuous-valued physiological parameters that are central to diagnosis. One significant challenge in this regard is the imbalanced nature of the dataset, as a low prevalence of abnormal conditions can lead to heavily skewed data that results in inaccurate predictions and a lack of certainty in such predictions when deployed. Addressing this, we propose a new framework, HypUC, for imbalanced probabilistic regression in medical time series. Our approach incorporates ideas from probabilistic machine learning for uncertainty estimation. We also introduce a new calibration method that provides reliable uncertainty estimates that generalize well to a diverse range of test sets. Additionally, we present a method for using these calibrated uncertainties to improve decision-making through an ensemble of gradient-boosted learners. Furthermore, we demonstrate an entropy-based technique to flag unreliable predictions. We evaluate our approach on a large, real-world dataset of ECGs collected from millions of patients with various medical conditions. Our approach outperforms several baselines while also providing calibrated uncertainty estimates for many diagnostic problems indicating its suitability for clinical use and real-world deployment.

1 Introduction

Electrocardiogram (ECG) signals are widely used in the diagnosis and management of cardiovascular diseases. However, manual analysis of ECG signals can be time-consuming and subject to human error. Automated analysis of ECG signals using machine learning (ML) techniques has the potential to improve the accuracy and efficiency of ECG signal analysis. Moreover, it will open up the avenue for remote patient monitoring for managing the cardiovascular health of the patients at scale. Recent advances in ML have led to the development of various algorithms for the automated analysis of ECG signals. These include methods for ECG signal classification and diagnosis of specific cardiovascular diseases Hannun et al. (2019). Convolutional neural networks (CNNs) and recurrent neural networks (RNNs) are commonly used for these tasks Siontis et al. (2021); Şentürk et al. (2018). Existing literature often focuses on classifying the medical time series (in one of the disease-related classes) Siontis et al. (2021); Hannun et al. (2019) instead of regressing the continuous-valued physiological parameter on which the diagnosis is based Von Bachmann et al. (2022). But often, solving the regression problem is more valuable than classification; for instance, consider the problem of measuring electrolytes like potassium in the blood. Excess of potassium (known as *Hyperkalemia*) can be fatal. Detecting if the patient has Hyperkalemia non-invasively

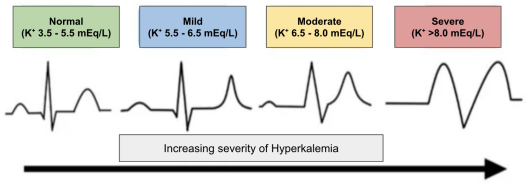

Figure 1: Severity of the Hyperkalemia will decide the diagnosis. Therefore binary classification is not sufficient.

by analyzing the ECG is desirable and has been studied previously Galloway et al. (2019). However, a binary classifier does not distinguish between the different severity of Hyperkalemia as shown in Figure 1 that may decide the final diagnosis for the condition, i.e., severe/mild hyperkalemia require different treatments.

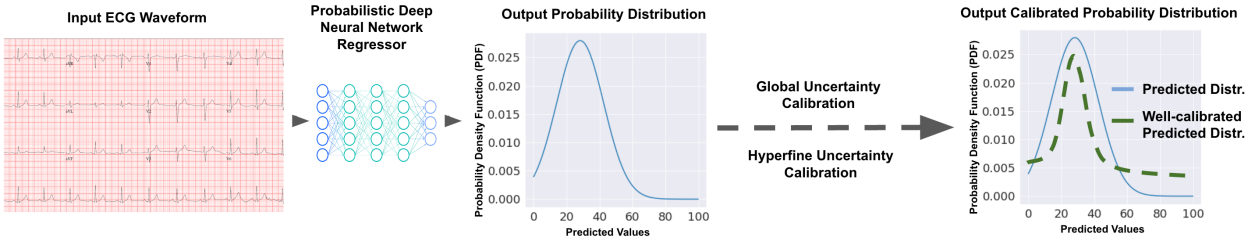

Figure 2: `HypUC`: The proposed framework processes the ECG waveform input to predict the parameters of the distribution (i.e., entire parametric distribution) representing the continuous value. The predicted distribution is subjected to global and hyperfine uncertainty calibration techniques to get a well-calibrated predictive distribution and uncertainty estimate as explained in Section 3.

One of the significant challenges in the real world for regressing the continuous-valued physiological parameter from medical time series is the imbalanced nature of the dataset, i.e., such datasets are often highly skewed if the prevalence of the abnormal condition is low. A heavily skewed dataset leads to a trained network that is inaccurate in the prediction and can not quantify uncertainty in predictions at deployment. To address these limitations, we propose a new framework called `HypUC`, shown in Figure 2. It leverages ideas from probabilistic regression and density estimation to tackle uncertainty-aware imbalanced regression with medical time series. We present a new and effective uncertainty calibration method (named hyperfine uncertainty calibration) that provides highly calibrated uncertainty estimates generalizing well to a large, diverse test set. Finally, we offer a new method that demonstrates how these calibrated uncertainties can be used algorithmically to improve decision-making consistently using an ensemble of gradient-boosted-based learners. We demonstrate the efficacy of our methods on a large real-world dataset of ECGs (collected using millions of patients with diverse medical conditions) to predict various physiological parameters such as *Left Ventricular Ejection Fraction*, *Serum Potassium Level*, *Age*, *Survival*, etc.

2 Related Work

In recent years, there has been a growing interest in using deep learning techniques for the analysis of electrocardiogram (ECG) signals. One of the main areas of focus has been the use of convolutional neural networks (CNNs) for ECG signal classification. A study by Rajpurkar et al. (2017); Hannun et al. (2019); Murugesan et al. (2018) used a CNN to classify ECG signals into normal and abnormal beats with high accuracy. Another study by Ebrahimi et al. (2020); Li et al. (2016) proposed a deep learning-based framework for arrhythmia detection using ECG signals. A study Han et al. (2022); Camps et al. (2018) proposed a deep learning-based method for predicting QRS complex locations in ECG signals. Another study Xu et al. (2019); Şentürk et al. (2018) presented a deep recurrent neural network (RNN) for predicting heart rate from ECG signals. Furthermore, there have been studies that have used deep learning for the diagnosis of specific cardiovascular diseases. The study Baloglu et al. (2019); Mahendran et al. (2021) proposed a deep learning-based method for detecting myocardial infarction (heart attack) using ECG signals. Another study Martinez-Ríos et al. (2021) presented a deep learning-based approach for identifying hypertension (high blood pressure) using ECG signals. The study Siontis et al. (2021); Feeny et al. (2020) proposed a deep learning-based model for the identification of atrial fibrillation using ECG signals, achieving high accuracy in detecting the condition. Another study Ko et al. (2020) used deep learning to identify ECG patterns associated with sudden cardiac death, demonstrating the potential for using these techniques for the prediction of cardiac events. Other works proposed deep learning algorithm for ECG-based risk stratification in patients with heart failure Bos et al. (2021). Previous works have also explored the use of deep learning in combination with other techniques, such as wavelet transforms. The study Feeny et al. (2020); Pant et al. (2022) proposed a deep learning-based framework for ECG signal analysis, which incorporated wavelet transforms for feature extraction and achieved high accuracy for arrhythmia detection. In recent times there

have also been many works on estimating uncertainties with DNNs Gal (2016); Lakshminarayanan et al. (2016); Blundell et al. (2015); Kendall & Gal (2017); Upadhyay et al. (2021b;a;c; 2022); Sudarshan et al. (2021); Daxberger et al. (2021); Laves et al. (2020b;a); Kumar et al. (2019); Kolczynski Jr et al. (2009); Garaud & Mallet (2011); Kuleshov et al. (2018); Xia et al. (2023); Jahmunah et al. (2023). However, they do not address imbalanced dataset (and the calibration of uncertainties in such setups).

3 Methodology

We first formulate the problem in Section 3.1. Preliminaries necessary to motivate the design choices for our framework (`HypUC`) are presented in Section 3.2.1 (on Imbalanced Regression), Section 3.2.2 (on Kernel Density Estimation), and Section 3.2.3 (on Uncertainty Estimation). In Section 3.3, we construct `HypUC`, that performs uncertainty-aware regression followed by a new calibration technique that significantly improves the quality of uncertainty estimates and utilizes a simple technique based on gradient-boosted decision trees to enhance decision-making using the predicted point estimate and well-calibrated uncertainty estimate.

3.1 Problem formulation

Let $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{N}$ be the training set with pairs from domain $\mathbf{X}$ and $\mathbf{Y}$ (i.e., $\mathbf{x}_i \in \mathbf{X}, \mathbf{y}_i \in \mathbf{Y}, \forall i$), where $\mathbf{X}, \mathbf{Y}$ lies in $\mathbb{R}^m$ and $\mathbb{R}$, respectively. While our proposed solution is valid for data of arbitrary input dimension, we present the formulation for medical time series (like ECGs) with applications that predict a desired physiological parameter helpful in making a diagnosis using ECG. Therefore, $(\mathbf{x}_i, \mathbf{y}_i)$ represents a pair of ECG and physiological quantity/parameter. For instance, to aid in diagnosing Hyperkalemia, $\mathbf{x}_i$ is ECG signal, and $\mathbf{y}_i$ is the serum potassium level (indicating the amount of potassium in the blood, which may be obtained from the blood test that serves as groundtruth labels for training).

We aim to learn a mapping, say $\mathbf{\Psi}(\cdot) : \mathbf{X} \rightarrow P(\mathbf{Y})$, that learns to map a given input ECG ($\mathbf{x}$) to a probability density function $P(\mathbf{y})$, indicating the probable values. We use a deep neural network (DNN) to parameterize $\mathbf{\Psi}$ with θ and solve an optimization problem to obtain optimal parameters θ^*, as described in Section 3.3.

3.2 Preliminaries

3.2.1 Imbalanced Regresison

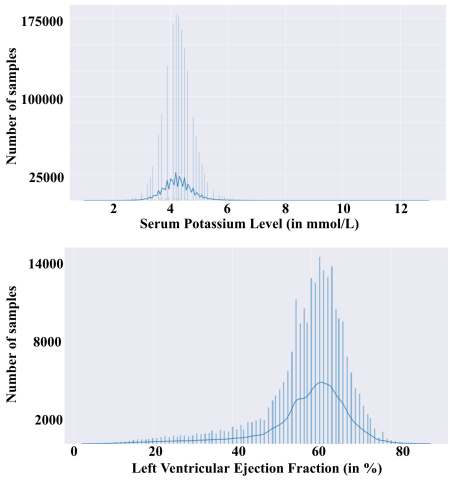

Figure 3: Imbalanced distribution of target values (top) Serum Potassium Levels and (bottom) Ejection Fraction in real-world medical datasets.

Imbalanced regression refers to the problem of training a machine learning model for regression on a dataset where the target variable is imbalanced, meaning that certain values appear more often than others, this is typical for medical datasets where different medical conditions appear with different prevalence. Figure 3 shows the density of Serum Potassium Levels and Left Ventricular Ejection fraction for a real-world dataset. We notice that most of the patients have serum potassium level in the range of 3.5 to 5.5, however, it is crucial for machine learning models to work well in the rare high potassium levels (defined above 5.5) that occurs rarely, i.e., the population density in the region is very low. Similar trend is observed for LVEF, where most of the data is concentrated in healthy range (between 40-60%) but low-LVEF regions (defined below 40%) which is important for model to identify, are sparsely present in the dataset. Imbalanced datasets can present challenges for training regression models, as the model may be biased towards the densely populated samples. The work in Yang et al. (2021); Steininger et al. (2021) proposes distribution smoothing for both labels and features, which explicitly acknowledge the effects of nearby targets and calibrate both label and learned feature distributions lead to better point estimates in sparsely populated regions, but does not tackle the problem of uncertainty estimation.

3.2.2 Kernel Density Estimation

In statistics, kernel density estimation (KDE) Terrell & Scott (1992); Chen (2017) is the application of kernel smoothing for probability density estimation. Given a set of i.i.d. samples $\{\mathbf{y}_1, \mathbf{y}_2 ... \mathbf{y}_n\}$ drawn from some unknown density function $\mathbf{f}$. The kernel density estimate provides an estimate of the shape of $\mathbf{f}$. This can be leveraged to train deep regression models with imbalanced dataset by estimating the density for each target value in the training set and use this density to create a weighting scheme to be applied with the loss for that sample. This approach can help to give more importance to samples with rare target values, which are often more difficult to predict. This is achieved by first estimating the density as:

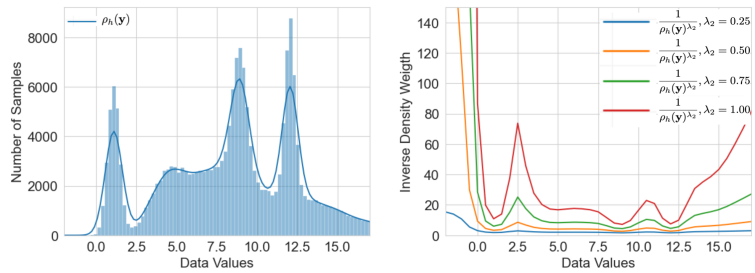

Figure 4: KDE and related weights for imbalanced targets. (left) Example of imbalanced target datasets with KDE. (right) Inverse of KDE with exponent used as weights for the loss term. Higher exponent leads to sharper weigths.

$$\rho_h(\mathbf{y}) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{K}_h(\mathbf{y} - \mathbf{y}_i) = \frac{1}{nh}\sum_{i=1}^{n} \mathcal{K}\left(\frac{\mathbf{y} - \mathbf{y}_i}{h}\right) \tag{1}$$

where $\mathcal{K}$ is the kernel — a non-negative function — and $h > 0$ is a smoothing parameter called the bandwidth. Scaled kernel $\mathcal{K}_h$ is defined as, $\mathcal{K}_h(\mathbf{y}) = \frac{1}{h}\mathcal{K}(\frac{\mathbf{y}}{h})$, and $\mathcal{K}$ is often set to standard normal density function. The weighting scheme can be designed by leveraging the fact that lower density samples must receive higher weights and vice-versa, a simple technique to achieve the same is to design the weight for target value $\mathbf{y}$ to be $\mathbf{w}(\mathbf{y}) = \frac{1}{\rho_h(\mathbf{y})^{\lambda_2}}$. Figure 4-(left) shows an example of imbalanced target distribution and Figure 4-(right) shows the designed weighting scheme, we note that hyperparameter λ_2 controls the sharpness of the weights.

3.2.3 Uncertainty Estimation

The works in Kendall & Gal (2017); Laves et al. (2020b); Upadhyay et al. (2022) highlight the key techniques to estimate the *irreducible* (i.e., aleatoric) uncertainty, and *reducible* (i.e., epistemic) uncertainty. As highlighted in Kendall & Gal (2017), capturing aleatoric uncertainty becomes crucial in the presence of large datasets. For an arbitrary DNN, $\mathbf{\Phi}(\cdot;\zeta) : \mathbf{X} \rightarrow \mathbf{Y}$, to estimate the aleatoric uncertainty, the model must estimate the parameters of $\mathcal{P}_{Y|X}$ which are then used to maximize the likelihood function. That is, for an input $\mathbf{x}_i$, the model produces a set of parameters representing the output given by, $\{\hat{\mathbf{y}}_i, \hat{\nu}_i \dots \hat{\rho}_i\} := \mathbf{\Phi}(\mathbf{x}_i;\zeta)$, that characterizes the distribution $\mathcal{P}_{Y|X}(\mathbf{y}; \{\hat{\mathbf{y}}_i, \hat{\nu}_i \dots \hat{\rho}_i\})$, such that $\mathbf{y}_i \sim \mathcal{P}_{Y|X}(\mathbf{y}; \{\hat{\mathbf{y}}_i, \hat{\nu}_i \dots \hat{\rho}_i\})$. The likelihood $\mathscr{L}(\zeta; \mathcal{D}) := \prod_{i=1}^{N} \mathcal{P}_{Y|X}(\mathbf{y}_i; \{\hat{\mathbf{y}}_i, \hat{\nu}_i \dots \hat{\rho}_i\})$ is then maximized to estimate the optimal parameters of the network. Moreover, the distribution $\mathcal{P}_{Y|X}$ is often chosen such that uncertainty can be estimated using a closed form solution $\mathscr{F}$ of estimated parameters from the DNN, i.e.,

$$\{\hat{\mathbf{y}}_i, \hat{\nu}_i \dots \hat{\rho}_i\} := \mathbf{\Phi}(\mathbf{x}_i;\zeta) \text{ with } \zeta^* := \underset{\zeta}{\operatorname{argmax}}\ \mathscr{L}(\zeta; \mathcal{D}) = \underset{\zeta}{\operatorname{argmax}} \prod_{i=1}^{N} \mathcal{P}_{Y|X}(\mathbf{y}_i; \{\hat{\mathbf{y}}_i, \hat{\nu}_i \dots \hat{\rho}_i\}) \tag{2}$$

$$\text{Uncertainty}(\hat{\mathbf{y}}_i) = \mathscr{F}(\hat{\nu}_i \dots \hat{\rho}_i) \tag{3}$$

It is common to use a *heteroscedastic* Gaussian distribution for $\mathcal{P}_{Y|X}$ Kendall & Gal (2017); Laves et al. (2020b), in which case $\mathbf{\Phi}(\cdot;\zeta)$ is designed to predict the *mean* and *variance* of the Gaussian distribution, i.e., $\{\hat{\mathbf{y}}_i, \hat{\sigma}_i^2\} := \mathbf{\Phi}(\mathbf{x}_i;\zeta)$, and the predicted *variance* itself can be treated as uncertainty in the prediction. The optimization problem becomes,

$$\zeta^* = \underset{\zeta}{\operatorname{argmax}} \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\hat{\sigma}_i^2}} e^{-\frac{|\hat{\mathbf{y}}_i - \mathbf{y}_i|^2}{2\hat{\sigma}_i^2}} = \underset{\zeta}{\operatorname{argmin}} \sum_{i=1}^{N} \frac{|\hat{\mathbf{y}}_i - \mathbf{y}_i|^2}{2\hat{\sigma}_i^2} + \frac{\log(\hat{\sigma}_i^2)}{2} \tag{4}$$

$$\text{Uncertainty}(\hat{\mathbf{y}}_i) = \hat{\sigma}_i^2. \tag{5}$$

3.3 Building HypUC

In this section, we explain the building blocks for our framework, which consists of three major steps: (i) Optimization for improved continuous-value point and uncertainty estimation, (ii) Post-training hyperfine uncertainty calibration, (iii) Improved decision-making with an ensemble of gradient-boosted learners.

3.3.1 Optimization for Improved Continuous-value Point and Uncertainty Estimation

As described in Section 3.1, we aim to learn the estimation function $\mathbf{\Psi}(\cdot;\theta)$. The DNN $\mathbf{\Psi}$ takes the medical time series as the input and predicts the probability density function parameters for the output. That is, $\mathbf{\Psi}(\mathbf{x};\theta) = \{\hat{\mathbf{y}}, \hat{\sigma}\} = \{[\mathbf{\Psi}(\mathbf{x}_i;\theta)]_{\hat{y}}, [\mathbf{\Psi}(\mathbf{x}_i;\theta)]_{\hat{\sigma}}\}$. The optimal parameters of the DNN, θ^*, are obtained by:

$$\theta^* = \underset{\theta}{\operatorname{argmin}} \frac{1}{N} \sum_{i=1}^{N} \left\{ \lambda_1 * \left(\frac{|\,([\mathbf{\Psi}(\mathbf{x}_i;\theta)]_{\hat{y}} - \mathbf{y}_i)\,|}{\rho_h(\mathbf{y}_i)^{\lambda_2}} \right) + \lambda_3 * \left(\frac{|\,([\mathbf{\Psi}(\mathbf{x}_i;\theta)]_{\hat{y}} - \mathbf{y}_i)\,|^2}{[\mathbf{\Psi}(\mathbf{x}_i;\theta)]^2_{\hat{\sigma}}} + \log\left([\mathbf{\Psi}(\mathbf{x}_i;\theta)]^2_{\hat{\sigma}}\right) \right) \right\}. \quad (6)$$

Where (i) $\rho(\mathbf{y})$ is the Gaussian kernel density estimate (fit on the training labels) for the target value $\mathbf{y}$, (ii) $[\mathbf{\Psi}(\mathbf{x}_i;\theta)]_{\hat{y}}$ is the $\hat{\mathbf{y}}_i$ estimate given by the network, (iii) $[\mathbf{\Psi}(\mathbf{x}_i;\theta)]_{\hat{\sigma}}$ is the $\hat{\sigma}_i$ estimate given by the network. $\lambda_1, \lambda_2, \lambda_3, h$ are the hyperparameters.

In the above optimization, the term $\left(\frac{|([\mathbf{\Psi}(\mathbf{x}_i;\theta)]_{\hat{y}} - \mathbf{y}_i)|}{\rho(\mathbf{y}_i)^{\lambda_2}}\right)$ computes the discrepancy between the continuous value prediction and the groundtruth, where the discrepancy is weighed down if the groundtruth label is present in abundance, i.e., density $\rho(\mathbf{y})$ is high. The discrepancy is weighed up if the label density is low, addressing the imbalanced nature of the dataset. The term, $\left(\frac{|([\mathbf{\Psi}(\mathbf{x}_i;\theta)]_{\hat{y}} - \mathbf{y}_i)|^2}{\mathbf{\Psi}(\mathbf{x}_i;\theta)]^2_{\hat{\sigma}}} + \log\left(\mathbf{\Psi}(\mathbf{x}_i;\theta)]^2_{\hat{\sigma}}\right)\right)$, in the above optimization indicates negative-log-likelihood of heteroscedastic Gaussian distribution assumed for each prediction. The hyperparameters λ_1, λ_3 control the relative contribution of the density-based discrepancy term and the heteroscedastic negative-log-likelihood term that learns to estimate the uncertainty.

3.3.2 Post-training Hyperfine Uncertainty Calibration

Once the DNN is trained, i.e., the optimal parameters θ^* have been found as described in the above optimization step, the DNN $\mathbf{\Psi}(;\theta^*)$ is capable of estimating both the target continuous-value and the uncertainty in the prediction. However, the uncertainty estimates obtained are biased and are not well calibrated Laves et al. (2020a); Bishop & Nasrabadi (2006); Hastie et al. (2009). Therefore, we propose a new technique to calibrate the uncertainty estimates. We first find a global optimal scaling factor s^* by solving the following optimization problem on the validation set $\mathcal{D}_{valid} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{M}$, where the parameters of the DNN are fixed to the optimal obtained above (i.e., θ^*) as described in Laves et al. (2020a),

$$s^* = \underset{s}{\operatorname{argmin}} \left(M \log(s) + \frac{1}{2s^2} \sum_{i=1}^{M} \frac{|\,([\mathbf{\Psi}(\mathbf{x}_i;\theta^*)]_{\hat{y}} - \mathbf{y}_i)\,|^2}{[\mathbf{\Psi}(\mathbf{x}_i;\theta^*)]^2_{\hat{\sigma}}} \right) \quad (7)$$

We then run the model in inference mode and generate the estimates for all the samples in the validation set, i.e., we get a new augmented validation set, $\mathcal{D}_{valid-aug} = \{\mathbf{x}_i, \mathbf{y}_i, \hat{\mathbf{y}}_i, \hat{\sigma}_i^2\}_{i=1}^{M}$. We then create hyperfine bins of uniform length spanning the entire range of predicted continuous-values, i.e., we create $b = int\left(\frac{\mathbf{y}_{max} - \mathbf{y}_{min}}{\delta}\right)$ bins, where $\delta << \mathbf{y}_{max} - \mathbf{y}_{min}$ leading to hyperfine bins, where $\mathbf{y}_{max} = max(\{\mathbf{y}_i\}_{i=1:M})$ and $\mathbf{y}_{min} = min(\{\mathbf{y}_i\}_{i=1:M})$. The nth bin being B_n spanning $[\mathbf{y}_{min} + (n-1)\delta, \mathbf{y}_{min} + n\delta]$. Each sample in the validation set then belongs to one of the bins. We then find a bin-wise scaling factor for each of the above hyperfine bins by solving an optimization problem. The bin-wise scaling factor for bin B_n, η_{B_n}, is given by,

$$\eta_{B_n} = \underset{\eta}{\operatorname{argmin}} \left(\frac{len(\{|\hat{\mathbf{y}}_i - \mathbf{y}_i| < \eta s^* \hat{\sigma}_i\}_{\hat{\mathbf{y}}_i \in B_n})}{len(B_n)} > \xi \right) \quad (8)$$

Where, $len(\cdot)$ is the function that gives the number of elements in the input set and ξ represents the minimum acceptable fraction in each bin where the discrepancy between the prediction and the groundtruth

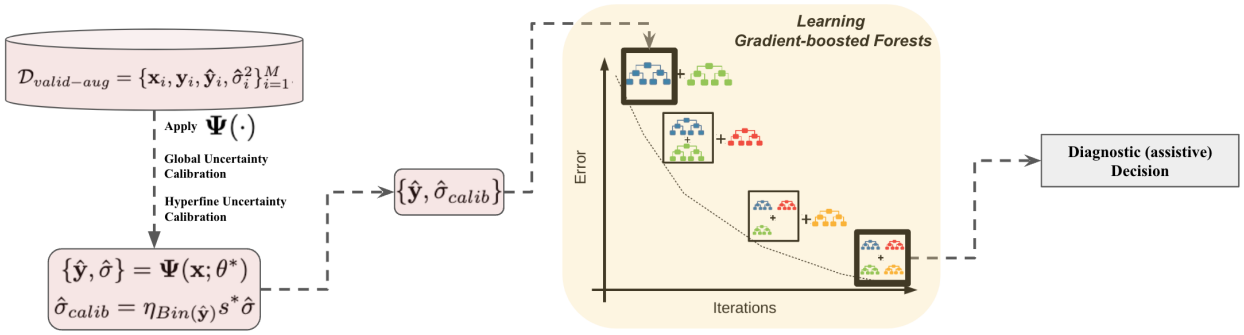

Figure 5: Gradient-boosted based ensemble of decision trees incorporating calibrated uncertainty (via both global and hyperfine calibration) and point estimate to help in diagnostic decision making.

(i.e., $|\hat{\mathbf{y}} - \mathbf{y}|$) is captured by the scaled uncertainty estimates (i.e., $\eta s^* \hat{\sigma}_i$). The above scaling factors allow us to compute the well calibrated uncertainty estimates for each prediction. Given input $\mathbf{x}$, the continuous value prediction ($\hat{\mathbf{y}}$) and the calibrated uncertainty ($\hat{\sigma}_{calib}$) are given by,

$$\{\hat{\mathbf{y}}, \hat{\sigma}\} = \mathbf{\Psi}(\mathbf{x}; \theta^*) \text{ and } \hat{\sigma}_{calib} = \eta_{Bin(\hat{\mathbf{y}})} s^* \hat{\sigma} \tag{9}$$

Where, $Bin(\hat{\mathbf{y}})$ is the hyperfine bin to which $\hat{\mathbf{y}}$ belongs $\eta_{Bin(\hat{\mathbf{y}})}$ is the bin-wise scaling factor of the bin.

3.3.3 Improved Decision-making with an Ensemble of Gradient-boosted Learners

Medical professionals often diagnose the condition based on continuous value lying in a certain standardized range Galloway et al. (2019); Bachtiger et al. (2022). We propose a technique that utilizes both the predicted continuous value and calibrated uncertainty estimate to infer the standardized range (that is better than relying on just the continuous-value point estimate). The proposed technique uses predictions (both $\hat{\mathbf{y}}$ and $\hat{\sigma}_{calib}$) from the trained DNN ($\mathbf{\Psi}$) as input features to an ensemble of decision trees that uses a gradient-boosting technique to learn a random forest (represented by $\mathbf{\Upsilon}$ with N_{trees} trees and d_{tree} as the maximum possible depth for each tree) on dataset $\mathcal{D}_{train} \cup \mathcal{D}_{valid}$, that infers the standardized-range/class in which the predicted continuous-value belongs, accounting for the uncertainty in the predictions,

$$\text{Class}(\mathbf{x}) = \mathbf{\Upsilon}\left([\mathbf{\Psi}(\mathbf{x}; \theta^*)]_{\hat{y}}, \eta_{Bin([\mathbf{\Psi}(\mathbf{x}; \theta^*)]_{\hat{y}})} s^* [\mathbf{\Psi}(\mathbf{x}; \theta^*)]_{\hat{\sigma}}\right). \tag{10}$$

4 Experiments and Results

4.1 Tasks and Datasets

Our dataset is curated at `DEIDENTIFIED-MEDICAL-SITE` and consists of over *2.5 million* patients, with over *8.9 million* 12-lead ECGs coming from different types of hardware, covering a wide range of medical conditions, age, and demographic variables. The data collection is from year 1980 to 2020. More details about the dataset and the collecting organization will be revealed post publication.

4.1.1 Estimating Age from ECG

Estimating age from an electrocardiogram (ECG) has gained increasing attention in recent years due to the potential for non-invasive age estimation to aid in diagnosing and treating age-related conditions. Previous works Attia et al. (2019) have shown that, using deep learning, it is possible to analyze the ECG signals such that the neural network can identify patterns in the ECG waveform that change with age, which can be used to estimate an individual's cardiovascular age. For our experiments, we curated a diverse dataset of 8.8 million ECGs (along with the patient's age). We used 6.7/0.75/1.35 million for train/val/test datasets.

4.1.2 Estimating Survival (time-to-death) from ECG

Estimating survival, or time to death, from an electrocardiogram (ECG) is a relatively new field of study that has gained momentum in recent years as it may aid in treatment and management of life-threatening

conditions Diller et al. (2019); Tripoliti et al. (2017). We curate the data to consist only of patients for which the death has been recorded. The time interval from ECG recording to death for such patients provide groundtruth for survival (time-to-death). In total there were 3 million ECGs with the corresponding time-to-death, we used 2.3/0.26/0.44 million for train/val/test sets.

4.1.3 Estimating Level of Potassium in Blood from ECG: Detecting Hyperkalemia

Serum potassium level is a measure of the amount of potassium in the blood. Potassium is an electrolyte that is essential for the proper functioning of the body's cells, tissues, and organs. It helps regulate the balance of fluids in the body, maintain normal blood pressure, and support proper muscle function, among other functions. The normal range for serum potassium levels is typically considered to be between 3.5-5.0 millimoles per liter (mmol/L). Levels below 3.5 mmol/L are considered to be low (*hypokalemia*) and levels above 5.0 mmol/L are considered to be high (*hyperkalemia*). Both low and high potassium levels can have serious health consequences. While a blood test is used to measure the serum potassium level, recent works have evaluated machine learning techniques to infer the potassium level using ECGs, this can play a critical role in remote patient monitoring Galloway et al. (2019; 2018); Feeny et al. (2020); Ahn et al. (2022). In total there are over 4 million ECGs with the corresponding serum potassium level (derived using blood test), we used 2.1/0.4/1.7 million for train/val/test sets.

4.1.4 Estimating Left Ventricular Ejection Fraction from ECG: Detecting low LVEF

Low *left ventricular ejection fraction* (LVEF) is a condition in which the left ventricle of the heart is unable to pump a sufficient amount of blood out of the heart and into the body. The left ventricle is the main pumping chamber of the heart and is responsible for pumping oxygenated blood to the rest of the body. LVEF is typically measured as a percentage, with a normal range being between $55 - 70\%$. A LVEF below 40% is generally considered to be low and may indicate heart failure. LVEF can be evaluated through various non-invasive tests, including an echocardiogram, a nuclear stress test. But ECG can still provide valuable information about heart function and can be used in conjunction with other tests to assess LVEF and diagnose heart conditions Noseworthy et al. (2020); Vaid et al. (2022). In our experiments, we leverage DNNs to infer the LVEF - this can help reduce the costs associated with overall diagnosis and treatment by reducing the need for more expensive Echocardiogram. In total there are over 0.6 million ECGs with the corresponding LVEF (derived using Echo), we used 0.3/0.1/0.2 million for train/val/test sets.

4.2 Compared Methods and Training details

To perform a comparative study with 12-lead ECG time series, we create a backbone DNN architecture adapting ResNets into a 1D version (with 1D convolutional layers) as discussed in previous works Attia et al. (2019). The DNN backbone arictecture is used to create class of (i) deterministic regression models, (ii) class of probabilistic regression models, and (iii) class of classification models for some of the tasks. For the probabilistic regression we append the backbone architecture with two smaller feed-forward heads to estimate the parameters of the predictive distribution as described in Kendall & Gal (2017); Upadhyay et al. (2022). The deterministic regression model is trained with (i) L2 loss, termed as `Regres.-L2`, (ii) L1 loss termed as `Regres.-L1`, (iii) We incorporate the density based weights to modulate the loss terms, we call this `Regres.-L1-KDEw`. We train two variants of probabilistic regression models (i) the standard *heteroscedastic Gaussian* model that estimates the mean and variance for the prediction and trained to maximize the likelihood as described in Von Bachmann et al. (2022), called `Regres.-w.-U` and (ii) Our proposed model `HypUC`, that incorporates a novel calibration technique along with gradient-boosted corrections. For serum potassium level estimation (and LVEF estimation), we also compare with a binary classifier that predicts whether the given ECG has Hyperkalemia (and low LVEF for LVEF estimation). We train all the models using Adam optimizer Kingma & Ba (2014), with (β_1, β_2) set to (0.9,0.999). The learning rate used to train models was $1e^{-4}$. For `HypUC`, the hyper-parameters $\{\lambda_1, h, \lambda_2, \lambda_3\}$ are set to (i) $\{1, 0.7, 0.8, 1e^{-3}\}$ for Age estimation. (ii) $\{1, 0.5, 0.5, 1e^{-4}\}$ for Survival estimation. (iii) $\{1, 1, 0.2, 1e^{-4}\}$ for Serum potassium estimation. (iv) $\{1, 1.5, 0.1, 1e^{-4}\}$ for Left Ventricular Ejection Fraction (LVEF) estimation.

4.3 Results

Table 1 shows the performance of various methods on different tasks and datasets. For all the regression tasks, we compute MSE and MAE to measure the discrepancy between the predictions and the groundtruth. Moreover, spearman and pearson correlation coefficient indicate the positive association between the predicted point estimates from the network and the groundtruth target variable. High correlation indicate that the prediction will be able to indicate the severity of the medical condition (which maybe helpful in prescribing the treatment as explained in Section 1). For probabilistic models, the quality of uncertainty estimates is measured by UCE and NLL Laves et al. (2020a); Upadhyay et al. (2022). Finally, for the binary classification tasks, we also report the AUC, Sensitivity, Specificity, Positive/Negative Predictive values (PPV/NPV).

4.3.1 Task Specific Performance

As indicated in Table 1, for *Survival Estimation*, the regression model trained with L2 loss (`Regres.-L2`) has a MSE of 100.12 (MAE of 8.49, Spearman/Pearson correlation of 0.30/0.30), which are worse than the performance of the model trained with L1 loss (`Regres.-L1`) (MSE of 73.21, MAE of 7.78, Spearman/Pearson correlation of 0.38/0.41), this is inline with previous works. Moreover, we could further improve the performance of the regression model trained with L1 loss by incorporating kernel density-based weights (`Regres.-L1-KDEw`), as described in Section 3.3, that accounts for imbalance in the dataset and leads to a performance with MSE of 67.97 (MAE of 6.30, Spearman/Pearson correlation of 0.51/0.53). Additionaly, we notice that heteroscedastic Gaussian probabilistic regression model (`Regres.-w.-U`) from Von Bachmann et al. (2022) performs better than homoscedastic Gaussian model (`Regres.-L2`) but not as good as (`Regres.-L1`). However, `Regres.-w.-U` estimates the uncertainty in the prediction (with UCE/NLL of 2.37/5.89), but our proposed `HypUC` not only improves regression performance (with MSE/MAE of 54.62/5.38) but the quality of uncertainty as well (with UCE/NLL of 0.58/3.45). We observe a similar trend for the *Age Estimation*.

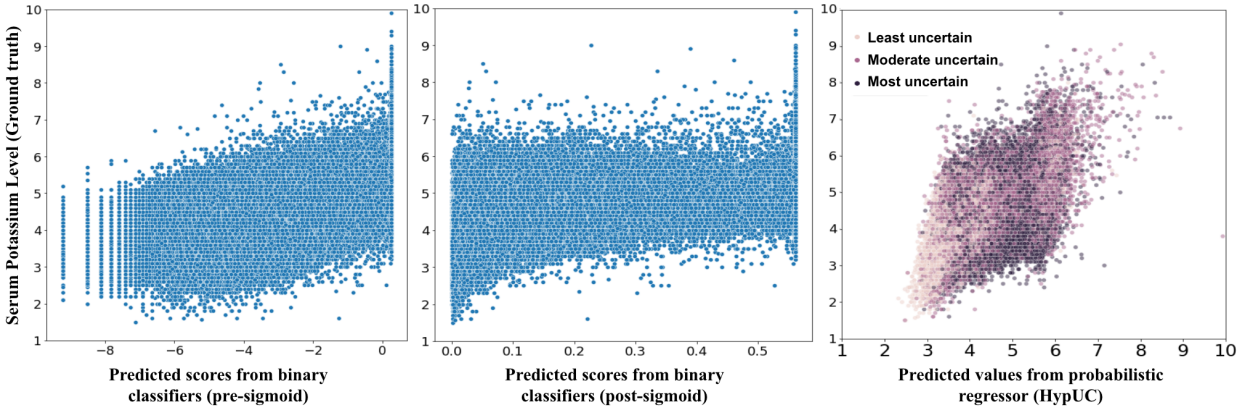

Figure 6: Serum Potassium Estimation model score correlation with groundtruth Serum Potassium levels. (left) Model scores (pre-activation) for the binary classifier. (middle) Model scores (post-activation) for the binary classifier. (right) Model scores for the proposed `HypUC`. The proposed `HypUC` not only produce better correlated outputs, but can also quantify the uncertainty estimates.

For *Serum Potassium Estimation*, we compared with a binary classifier baseline (`Binary-Classifier` in Table 1), an approach found in Galloway et al. (2019), which predicts if a given ECG corresponds to Hyperkalemia or not, instead of estimating the serum potassium level. While such a binary classifier achieves a good classification performance as indicated by an AUC of 0.87, it does not gauge the severity of the condition that may decide the diagnosis. Figure 6 shows the correlation between the ground truth serum potassium level compared to the scores derived from binary classifier (Figure 6-left and middle), the lack of high positive correlation (Spearman correlation of 0.39) indicates that scores from the model can not be used as an indication for the severity of the medical condition (within Hyperkalemia), but can be used to discriminate Hyperkalemia from non-Hyperkalemia cases. We also train different kinds of regression models to regress the serum potassium level. As indicated in Table 1, we see that regression models generally produce outputs that are better correlated with the groundtruth (e.g., `Regres.L2` has a Spearman correlation coef-

ficient of 0.51 vs. 0.39 for `Binary Classifier`). Moreover, we observe that among deterministic regression models `Regres.-L1-KDEw` performs the best with MSE of 0.19 (MAE of 0.33). The probabilistic regressor `Regres.-w.-U` performs better than `Regres.-L2` and provides uncertainty estimates as well (UCE/NLL of 1.83/8.76). But, `HypUC` provides the best regression performance along with better uncertainty estimates. Additionally, the outputs from the regression models can be used to classify the ECG into Hyperkalemia or not (similar to binary classifier), we notice that all the regression models perform comparable classification to classifier. In particular, `HypUC` has AUC of 0.89 which is higher than the binary classifier (AUC of 0.87).

Tasks	Methods	MSE	MAE	Spear. Corr.	Pears. Corr.	UCE	NLL	AUC	Sensiti.	Specifi.	PPV	NPV
Survival Estimation	`Regres.-L2`	100.12	8.49	0.30	0.30	NA	NA	NA	NA	NA	NA	NA
	`Regres.-L1`	73.21	7.78	0.38	0.41	NA	NA	NA	NA	NA	NA	NA
	`Regres.-L1-KDEw`	67.97	6.30	0.51	0.53	NA	NA	NA	NA	NA	NA	NA
	`Regres.-w.-U`	84.62	8.13	0.35	0.38	2.37	5.89	NA	NA	NA	NA	NA
	`HypUC`	**54.62**	**5.38**	**0.56**	**0.61**	**0.58**	**3.45**	NA	NA	NA	NA	NA
Age Estimation	`Regres.-L2`	151.69	9.47	0.58	0.60	NA	NA	NA	NA	NA	NA	NA
	`Regres.-L1`	116.52	7.93	0.73	0.75	NA	NA	NA	NA	NA	NA	NA
	`Regres.-L1-KDEw`	99.77	7.64	0.78	0.81	NA	NA	NA	NA	NA	NA	NA
	`Regres.-w.-U`	136.27	8.22	0.68	0.72	13.32	7.84	NA	NA	NA	NA	NA
	`HypUC`	**74.9**	**6.70**	**0.84**	**0.88**	**1.06**	**3.70**	NA	NA	NA	NA	NA
Serum Potassium Estimation	`Binary-Classifier`	NA	NA	0.39	0.41	NA	NA	0.87	0.74	0.85	0.10	0.99
	`Regres.-L2`	0.26	0.52	0.51	0.54	NA	NA	0.84	0.70	0.86	0.10	0.99
	`Regres.-L1`	0.19	0.34	0.52	0.54	NA	NA	0.86	0.70	0.87	0.09	0.99
	`Regres.-L1-KDEw`	0.19	0.33	0.52	0.55	NA	NA	0.86	0.72	0.85	0.11	0.99
	`Regres.-w.-U`	0.23	0.38	0.51	0.55	1.83	8.76	0.86	0.71	0.86	0.10	0.99
	`HypUC`	0.20	**0.32**	**0.52**	**0.55**	**0.41**	**3.12**	**0.89**	**0.77**	0.86	0.10	**0.99**
Ejection Fraction Estimation	`Binary-Classifier`	NA	NA	0.28	0.32	NA	NA	0.91	0.77	0.87	0.09	0.98
	`Regres.-L2`	171.37	13.56	0.37	0.39	NA	NA	0.90	0.75	0.84	0.09	0.98
	`Regres.-L1`	153.66	11.78	0.41	0.42	NA	NA	0.91	0.76	0.86	0.10	0.98
	`Regres.-L1-KDEw`	141.62	10.71	0.41	0.43	NA	NA	0.91	0.75	0.87	0.09	0.99
	`Regres.-w.-U`	165.81	12.24	0.41	0.44	10.56	9.83	0.90	0.74	0.85	0.10	0.99
	`HypUC`	**133.26**	**10.21**	**0.43**	**0.45**	**1.68**	**4.28**	**0.93**	**0.77**	**0.87**	**0.10**	**0.99**

Table 1: Quantitative results on 4 different tasks (Survival, Age, Serum Potassium, and Left Ventricular Ejection Fraction Estimation) showing the performance of various deterministic and probabilistic regression and classification models in terms of Mean Squared Error (MSE), Mean Absolute Error (MAE), Spearman Correlation (Spear. Corr.), Pearson Correlation (Pear. Corr.), Uncertainty Calibration Error (UCE), Negative Log-likelihood (NLL), AUC, Sensitivity, Specificity, Positive/Negative Predictive Values PPV/NPV. If for a task/method certain metric is not applicable we show NA.

A similar trend is observed for *Left Ventricular Ejection Fraction (LVEF) Estimation* where the proposed `HypUC` performs well in regression along with better calibrated uncertainty estimates. Moreover, when used as a classifier, it also yields a better performance than the binary classifier trained from scratch, emphasizing the benefits of tackling this problem with the proposed probabilistic regression framework, `HypUC`.

4.3.2 Calibration of Uncertainty Estimates on the Test-set

Figure 7 shows the groundtruth and the predicted distribution for Age/Survival/Serum Potassium/LVEF estimation tasks on the test dataset. The x-axis in the plot indicates the predicted point estimates from the `HypUC`. For a given point on x-axis: (i) Red bold curve shows the peak of the groundtruth distribution (i.e., the distribution of groundtruth values that correspond to predicted value given by the x-axis). (ii) Shaded

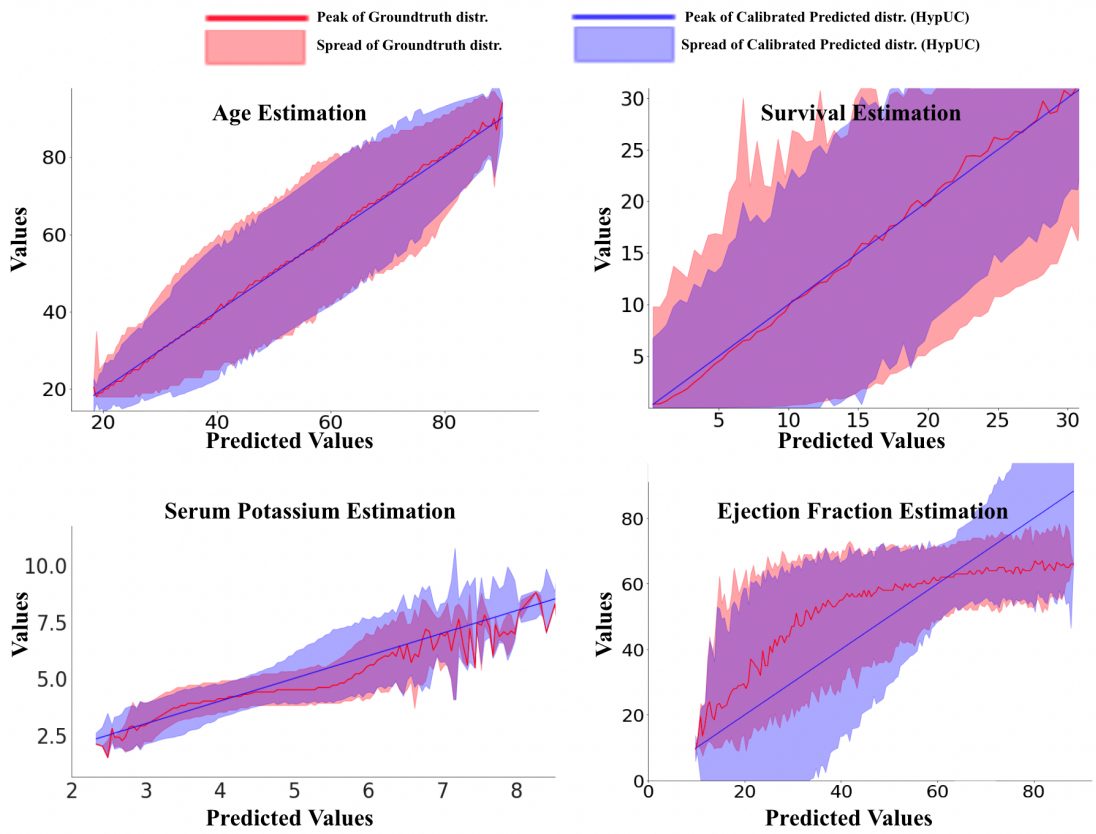

Figure 7: Groundtruth and predicted distributions for 4 different tasks (Age/Survival/Serum Potassium/LVEF estimation). Predicted distribution from `HypUC` tightly encompasses groundtruth distribution (compared to predicted distribution from `Regres.-w.-U`) indicating predictions are well-calibrated.

red region indicate the spread of the groundtruth distribution (i.e., 2 and 98 percentile of the groundtruth distribution). (iii) Blue bold curve shows the peak of the `HypUC`-predicted distribution. (iv) Blue shaded region indicate the spread of the groundtruth distribution (i.e., 2 and 98 percentile of the groundtruth distribution). We notice that predicted distribution from `HypUC` tightly encompasses groundtruth distribution (compared to predicted distribution from `Regres.-w.-U`) indicating predictions are well-calibrated. We calibrate `HypUC` using small validation sets but it generalizes well to much larger test sets.

4.3.3 Entropy-based Filtering to Trigger Human Expert Intervention

As the proposed `HypUC` produces well calibrated predictive distribution, we propose a technique to use the predicted distribution to flag the unreliable predictions. For a given input sample $\mathbf{x}$, the trained DNN produces $\{\hat{\mathbf{y}}, \hat{\sigma}\} = \mathbf{\Psi}(\mathbf{x}; \theta^*)$. We then apply the post-hoc calibration techniques to obtain $\{\hat{\mathbf{y}}, \hat{\sigma}_{calib}\}$. We then compute the entropy of the predicted distribution, i.e., $\mathcal{N}(\hat{\mathbf{y}}, \hat{\sigma}^2_{calib})$ given by $\mathcal{H}(\mathbf{z})$, i.e.,

$$\mathcal{H}(\mathbf{z}) = \int -p(\mathbf{z}) \log p(\mathbf{z}) \text{ , where } \mathbf{z} \sim \mathcal{N}(\hat{\mathbf{y}}, \hat{\sigma}^2_{calib}) \tag{11}$$

$$\mathcal{H}(\mathbf{z}) = -\mathbb{E}(\log \mathcal{N}(\hat{\mathbf{y}}, \hat{\sigma}^2_{calib})) = \log \sqrt{2\pi e \hat{\sigma}^2_{calib}} \tag{12}$$

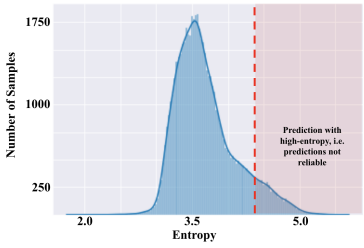

Figure 8: Distribution of entropy for LVEF validation set with a threshold of 10%-tile.

At the test time, if for the sample $\mathbf{x}$ the entropy $\log \sqrt{2\pi e \hat{\sigma}^2_{calib}} > \tau_{q,val}$ then the prediction is flagged as unreliable. The threshold $\tau_{q,val}$ is computed using the validation set as the $q-$%tile value from the distribution of entropy values as shown in Figure 8 for LVEF estimation using `HypUC`. Figure 9 evaluates the regression performance of `HypUC` (in terms

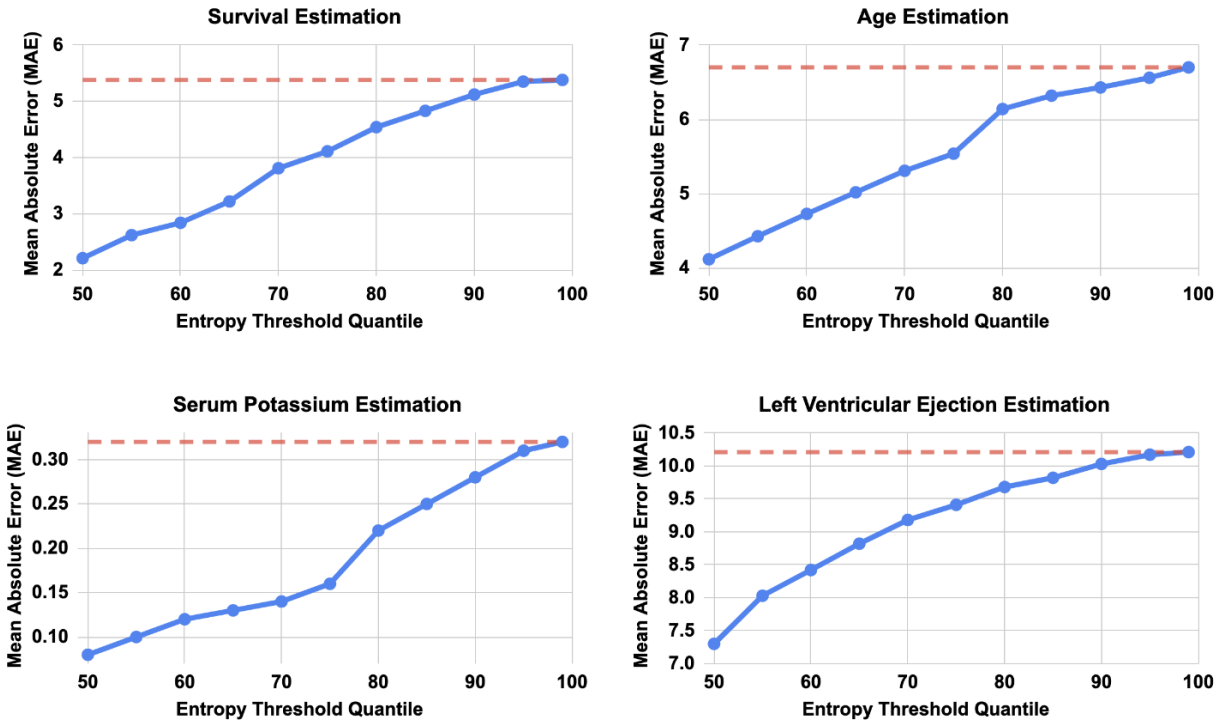

Figure 9: Evaluating `HypUC` on test-set by removing the unreliable (i.e., high entropy) predictions. We see consistent improvement as we decrease the entropy threshold based on quantile as described in Section 4.3.3

of MAE) on the test-sets for various tasks by removing the unreliable predictions based on different thresholds derived from the validation set. We observe that as we remove the higher fraction of unreliable samples (i.e., decreasing the threshold $\tau_{q,val}$ by lowering the value of q) the performance improves. This phenomena demonstrates that we could enhance the predictions made by model by only showing reliable predictions and flagging the unreliable predictions (based on entropy measure) to be evaluated by human experts which may be critical for deploying machine learning models for healthcare applications in real-world.

5 Discussion and Conclusion

Automated analysis of medical time series (like ECG) using machine learning (ML) techniques open several avenues in large scale remote healthcare assistance while making more efficient use of existing procedures. Previous research in this domain has largely focused on classifying medical time series into one of the several disease classes rather than attempting to regress the continuous-valued physiological parameters that are central to diagnosis of many medical conditions. This is often because imbalanced dataset present in these domains make it challenging for ML models to be accurate and certain when deployed. In this work, we build a new method, `HypUC`, in an attempt to solve the imbalanced regression problems with medical time series. `HypUC` incorporates ideas from probabilistic machine learning to also perform uncertainty estimation. We also introduced a new hyperfine uncertainty calibration technique that provides reliable uncertainty estimates for a diverse range of test sets. Additionally, we present a method for using these calibrated uncertainties to improve decision-making through an ensemble of gradient-boosted learners. In our studies, we use the well-calibrated predictive distribution derived from `HypUC` in an entropy-based technique to flag the unreliable predictions at the inference time for human expert evaluation and improve the performance of the ML model. Our approach is demonstrated on a large, real-world dataset of ECGs collected from millions of patients with various medical conditions, and is shown to outperform several baselines while also providing calibrated uncertainty estimates suitable for clinical use and real-world deployment.

References

Joseph C Ahn, Zachi I Attia, Puru Rattan, Aidan F Mullan, Seth Buryska, Alina M Allen, Patrick S Kamath, Paul A Friedman, Vijay H Shah, Peter A Noseworthy, et al. Development of the ai-cirrhosis-ecg score: An electrocardiogram-based deep learning model in cirrhosis. *The American Journal of Gastroenterology*, 117(3):424–432, 2022.

Zachi I Attia, Paul A Friedman, Peter A Noseworthy, Francisco Lopez-Jimenez, Dorothy J Ladewig, Gaurav Satam, Patricia A Pellikka, Thomas M Munger, Samuel J Asirvatham, Christopher G Scott, et al. Age and sex estimation using artificial intelligence from standard 12-lead ecgs. *Circulation: Arrhythmia and Electrophysiology*, 12(9):e007284, 2019.

Patrik Bachtiger, Camille F Petri, Francesca E Scott, Se Ri Park, Mihir A Kelshiker, Harpreet K Sahemey, Bianca Dumea, Regine Alquero, Pritpal S Padam, Isobel R Hatrick, et al. Point-of-care screening for heart failure with reduced ejection fraction using artificial intelligence during ecg-enabled stethoscope examination in london, uk: a prospective, observational, multicentre study. *The Lancet Digital Health*, 4 (2):e117–e125, 2022.

Ulas Baran Baloglu, Muhammed Talo, Ozal Yildirim, Ru San Tan, and U Rajendra Acharya. Classification of myocardial infarction with multi-lead ecg signals and deep cnn. *Pattern Recognition Letters*, 122:23–30, 2019.

Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *ICML*, 2015.

J Martijn Bos, Zachi I Attia, David E Albert, Peter A Noseworthy, Paul A Friedman, and Michael J Ackerman. Use of artificial intelligence and deep neural networks in evaluation of patients with electrocardiographically concealed long qt syndrome from the surface 12-lead electrocardiogram. *JAMA cardiology*, 6(5):532–538, 2021.

Julià Camps, Blanca Rodríguez, and Ana Mincholé. Deep learning based qrs multilead delineator in electrocardiogram signals. In *2018 Computing in Cardiology Conference (CinC)*, volume 45, pp. 1–4. IEEE, 2018.

Yen-Chi Chen. A tutorial on kernel density estimation and recent advances. *Biostatistics & Epidemiology*, 1(1):161–187, 2017.

Erik Daxberger, Agustinus Kristiadi, Alexander Immer, Runa Eschenhagen, Matthias Bauer, and Philipp Hennig. Laplace redux-effortless bayesian deep learning. *NeurIPS*, 2021.

Gerhard-Paul Diller, Aleksander Kempny, Sonya V Babu-Narayan, Marthe Henrichs, Margarita Brida, Anselm Uebing, Astrid E Lammers, Helmut Baumgartner, Wei Li, Stephen J Wort, et al. Machine learning algorithms estimating prognosis and guiding therapy in adult congenital heart disease: data from a single tertiary centre including 10 019 patients. *European heart journal*, 40(13):1069–1077, 2019.

Zahra Ebrahimi, Mohammad Loni, Masoud Daneshtalab, and Arash Gharehbaghi. A review on deep learning methods for ecg arrhythmia classification. *Expert Systems with Applications: X*, 7:100033, 2020.

Albert K Feeny, Mina K Chung, Anant Madabhushi, Zachi I Attia, Maja Cikes, Marjan Firouznia, Paul A Friedman, Matthew M Kalscheur, Suraj Kapa, Sanjiv M Narayan, et al. Artificial intelligence and machine learning in arrhythmias and cardiac electrophysiology. *Circulation: Arrhythmia and Electrophysiology*, 13 (8):e007952, 2020.

Yarin Gal. Uncertainty in deep learning. 2016.

Conner D Galloway, Alexander V Valys, Frank L Petterson, Vivek P Gundotra, Daniel L Treiman, David E Albert, John J Dillon, Zachi I Attia, and Paul A Friedman. Non-invasive detection of hyperkalemia with a smartphone electrocardiogram and artificial intelligence. *Journal of the American College of Cardiology*, 71(11S):A272–A272, 2018.

Conner D Galloway, Alexander V Valys, Jacqueline B Shreibati, Daniel L Treiman, Frank L Petterson, Vivek P Gundotra, David E Albert, Zachi I Attia, Rickey E Carter, Samuel J Asirvatham, et al. Development and validation of a deep-learning model to screen for hyperkalemia from the electrocardiogram. *JAMA cardiology*, 4(5):428–436, 2019.

Damien Garaud and Vivien Mallet. Automatic calibration of an ensemble for uncertainty estimation and probabilistic forecast: Application to air quality. *Journal of Geophysical Research: Atmospheres*, 2011.

Chuang Han, Wenge Que, Songwei Wang, Jinying Zhang, Jie Zhao, and Li Shi. Qrs complexes and t waves localization in multi-lead ecg signals based on deep learning and electrophysiology knowledge. *Expert Systems with Applications*, 199:117187, 2022.

Awni Y Hannun, Pranav Rajpurkar, Masoumeh Haghpanahi, Geoffrey H Tison, Codie Bourn, Mintu P Turakhia, and Andrew Y Ng. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature medicine*, 25(1):65–69, 2019.

Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

V Jahmunah, EYK Ng, Ru-San Tan, Shu Lih Oh, and U Rajendra Acharya. Uncertainty quantification in densenet model using myocardial infarction ecg signals. *Computer Methods and Programs in Biomedicine*, 229:107308, 2023.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *NIPS*, 2017.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Wei-Yin Ko, Konstantinos C Siontis, Zachi I Attia, Rickey E Carter, Suraj Kapa, Steve R Ommen, Steven J Demuth, Michael J Ackerman, Bernard J Gersh, Adelaide M Arruda-Olson, et al. Detection of hypertrophic cardiomyopathy using a convolutional neural network-enabled electrocardiogram. *Journal of the American College of Cardiology*, 75(7):722–733, 2020.

Walter C Kolczynski Jr, David R Stauffer, Sue Ellen Haupt, and Aijun Deng. Ensemble variance calibration for representing meteorological uncertainty for atmospheric transport and dispersion modeling. *Journal of applied meteorology and climatology*, 2009.

Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. Accurate uncertainties for deep learning using calibrated regression. In *International conference on machine learning*, pp. 2796–2804. PMLR, 2018.

Ananya Kumar, Percy S Liang, and Tengyu Ma. Verified uncertainty calibration. *Advances in Neural Information Processing Systems*, 32, 2019.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *arXiv preprint arXiv:1612.01474*, 2016.

Max-Heinrich Laves, Sontje Ihler, Jacob F Fast, Lüder A Kahrs, and Tobias Ortmaier. Well-calibrated regression uncertainty in medical imaging with deep learning. In *Medical Imaging with Deep Learning*, pp. 393–412. PMLR, 2020a.

Max-Heinrich Laves, Sontje Ihler, Karl-Philipp Kortmann, and Tobias Ortmaier. Calibration of model uncertainty for dropout variational inference. *arXiv preprint arXiv:2006.11584*, 2020b.

Hongqiang Li, Danyang Yuan, Youxi Wang, Dianyin Cui, and Lu Cao. Arrhythmia classification based on multi-domain feature extraction for an ecg recognition system. *Sensors*, 16(10):1744, 2016.

Rakesh Kumar Mahendran, V Prabhu, V Parthasarathy, and A Mary Judith. Deep learning based adaptive recurrent neural network for detection of myocardial infarction. *Journal of Medical Imaging and Health Informatics*, 11(12):3044–3053, 2021.

Erick Martinez-Ríos, Luis Montesinos, Mariel Alfaro-Ponce, and Leandro Pecchia. A review of machine learning in hypertension detection and blood pressure estimation based on clinical and physiological data. *Biomedical Signal Processing and Control*, 68:102813, 2021.

Balamurali Murugesan, Vignesh Ravichandran, Keerthi Ram, SP Preejith, Jayaraj Joseph, Sharath M Shankaranarayana, and Mohanasankar Sivaprakasam. Ecgnet: Deep network for arrhythmia classification. In *2018 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, pp. 1–6. IEEE, 2018.

Peter A Noseworthy, Zachi I Attia, LaPrincess C Brewer, Sharonne N Hayes, Xiaoxi Yao, Suraj Kapa, Paul A Friedman, and Francisco Lopez-Jimenez. Assessing and mitigating bias in medical artificial intelligence: the effects of race and ethnicity on a deep learning model for ecg analysis. *Circulation: Arrhythmia and Electrophysiology*, 13(3):e007988, 2020.

Himanshu Pant, Hitesh Kumar Dhanda, and Sachin Taran. Sleep apnea detection using electrocardiogram signal input to fawt and optimize ensemble classifier. *Measurement*, 189:110485, 2022.

Pranav Rajpurkar, Awni Y Hannun, Masoumeh Haghpanahi, Codie Bourn, and Andrew Y Ng. Cardiologist-level arrhythmia detection with convolutional neural networks. *arXiv preprint arXiv:1707.01836*, 2017.

Ümit Şentürk, İbrahim Yücedağ, and Kemal Polat. Repetitive neural network (rnn) based blood pressure estimation using ppg and ecg signals. In *2018 2Nd international symposium on multidisciplinary studies and innovative technologies (ISMSIT)*, pp. 1–4. Ieee, 2018.

Konstantinos C Siontis, Peter A Noseworthy, Zachi I Attia, and Paul A Friedman. Artificial intelligence-enhanced electrocardiography in cardiovascular disease management. *Nature Reviews Cardiology*, 18(7): 465–478, 2021.

Michael Steininger, Konstantin Kobs, Padraig Davidson, Anna Krause, and Andreas Hotho. Density-based weighting for imbalanced regression. *Machine Learning*, 110(8):2187–2211, 2021.

Viswanath P Sudarshan, Uddeshya Upadhyay, Gary F Egan, Zhaolin Chen, and Suyash P Awate. Towards lower-dose pet using physics-based uncertainty-aware multimodal learning with robustness to out-of-distribution data. *Medical Image Analysis*, 2021.

George R Terrell and David W Scott. Variable kernel density estimation. *The Annals of Statistics*, pp. 1236–1265, 1992.

Evanthia E Tripoliti, Theofilos G Papadopoulos, Georgia S Karanasiou, Katerina K Naka, and Dimitrios I Fotiadis. Heart failure: diagnosis, severity estimation and prediction of adverse events through machine learning techniques. *Computational and structural biotechnology journal*, 15:26–47, 2017.

Uddeshya Upadhyay, Yanbei Chen, and Zeynep Akata. Robustness via uncertainty-aware cycle consistency. *NeurIPS*, 2021a.

Uddeshya Upadhyay, Yanbei Chen, Tobias Hepp, Sergios Gatidis, and Zeynep Akata. Uncertainty-guided progressive gans for medical image translation. In *MICCAI*, 2021b.

Uddeshya Upadhyay, Viswanath P Sudarshan, and Suyash P Awate. Uncertainty-aware gan with adaptive loss for robust mri image enhancement. In *IEEE ICCV Workshop*, 2021c.

Uddeshya Upadhyay, Shyamgopal Karthik, Yanbei Chen, Massimiliano Mancini, and Zeynep Akata. Bayescap: Bayesian identity cap for calibrated uncertainty in frozen neural networks. In *European Conference on Computer Vision*, pp. 299–317. Springer, 2022.

Akhil Vaid, Kipp W Johnson, Marcus A Badgeley, Sulaiman S Somani, Mesude Bicak, Isotta Landi, Adam Russak, Shan Zhao, Matthew A Levin, Robert S Freeman, et al. Using deep-learning algorithms to simultaneously identify right and left ventricular dysfunction from the electrocardiogram. *Cardiovascular Imaging*, 15(3):395–410, 2022.

Philipp Von Bachmann, Daniel Gedon, Fredrik K Gustafsson, Antônio H Ribeiro, Erik Lampa, Stefan Gustafsson, Johan Sundström, and Thomas B Schön. Ecg-based electrolyte prediction: Evaluating regression and probabilistic methods. *arXiv preprint arXiv:2212.13890*, 2022.

Tong Xia, Jing Han, and Cecilia Mascolo. Benchmarking uncertainty quantification on biosignal classification tasks under dataset shift. In *Multimodal AI in Healthcare*, pp. 347–359. Springer, 2023.

Ke Xu, Xinyu Jiang, Haoran Ren, Xiangyu Liu, and Wei Chen. Deep recurrent neural network for extracting pulse rate variability from photoplethysmography during strenuous physical exercise. In *2019 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, pp. 1–4. IEEE, 2019.

Yuzhe Yang, Kaiwen Zha, Ying-Cong Chen, Hao Wang, and Dina Katabi. Delving into deep imbalanced regression. In *International Conference on Machine Learning (ICML)*, 2021.