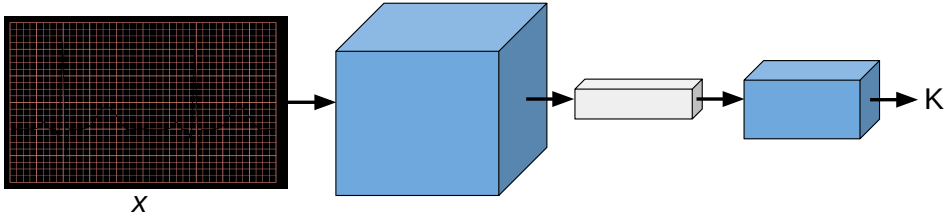
UPPSALA
UNIVERSITET

Can You Trust Your Regression Model's Uncertainty Under Distribution Shifts?

Fredrik K. Gustafsson
Uppsala University

SysCon μ seminar
September 15, 2022

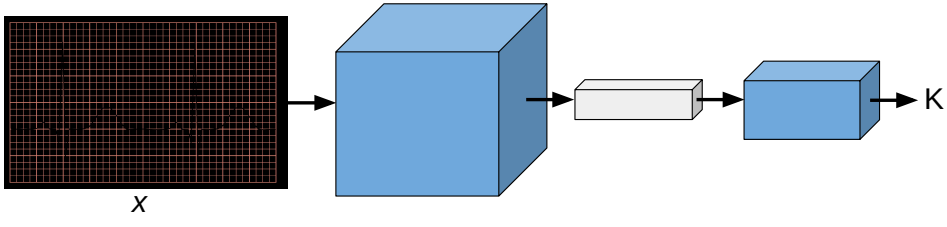
Motivating Example: ECG-Based Regression of Potassium Levels



Abnormal potassium (K^+) concentration levels in the human body can lead to serious heart conditions. The concentration level can be accurately measured via blood samples, but these are invasive and require relatively time-consuming analysis.

If the concentration level could be accurately monitored in real-time using an ECG-based regression model, potentially life-threatening conditions could be avoided.

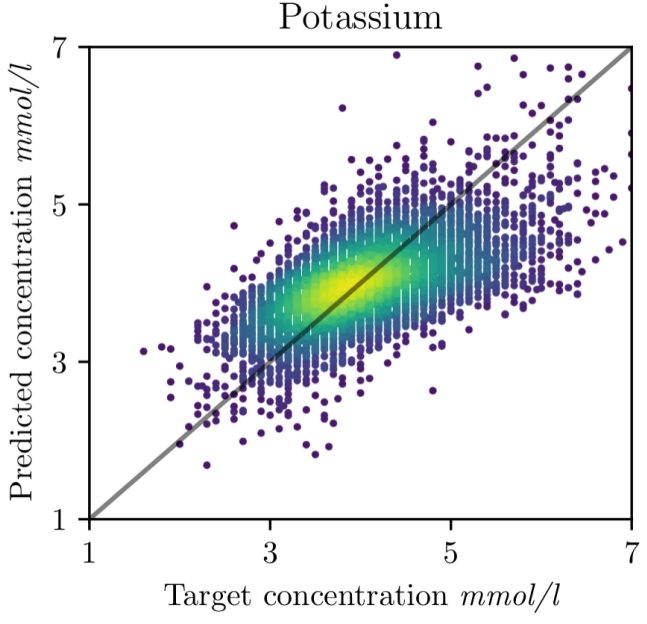
Motivating Example: ECG-Based Regression of Potassium Levels



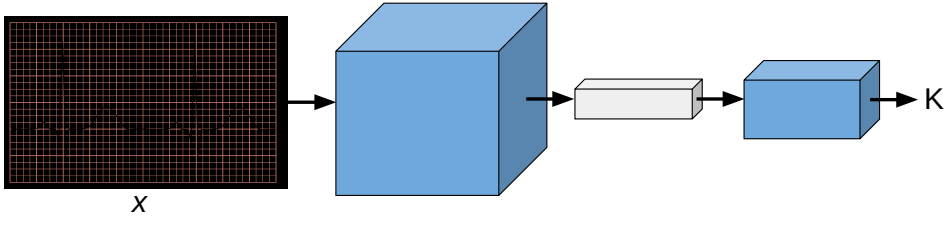
We recently trained a DNN on this task and obtained decent regression accuracy. To train the model, we utilized a large-scale dataset of over 290 000 ECGs from adult patients attending emergency departments at Swedish hospitals.

The ground truth regression targets were extracted from analyzed blood samples, collected within 60 *minutes of the corresponding ECGs.*

Motivating Example: ECG-Based Regression of Potassium Levels



Motivating Example: ECG-Based Regression of Potassium Levels



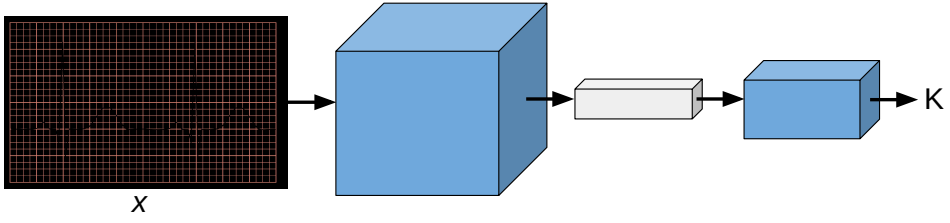
Could this model actually be employed in clinical practice at the university hospital?

No. The model fails to provide any measure of **uncertainty** for the predictions.

What if we explicitly model $p(y|x)$ with a Gaussian distribution $N\left(y; \mu_\theta(x), \sigma^2_\theta(x)\right)$ and use the variance $\sigma^2_\theta(x)$ as an estimate of uncertainty?

No. The model is often **overconfident** and outputs highly confident (small $\sigma^2_\theta(x)$) yet highly incorrect predictions, which arguably is even more dangerous than before.

Motivating Example: ECG-Based Regression of Potassium Levels



To be employed in a safety-critical medical application like this, the regression model must at least be **well calibrated**. If it outputs a prediction and e.g. a 90% prediction interval for each input, (at least) 90% interval coverage should actually be achieved.

The model must also remain well-calibrated under the wide range of **distribution shifts** which might be encountered during employment in clinical practice.

For example, if the model is trained on data collected solely from male patients in the year 2020, it must output well-calibrated predictions also for female patients in 2022.

Datasets

We collected 8 publicly available datasets for different image-based regression tasks, with various types of distribution shifts (*e.g., train on satellite images captured in densely populated American cities - test on images captured in a rural European area*).

2 synthetic datasets, 6 real-world datasets.

6 592 - 20 614 train images.

1 000 - 6 116 val images.

1 276 - 6 252 test images.

For convenience, we resized all images to size 64 64.

(some of the) Compared Methods

Conformal prediction.

Ensemble.

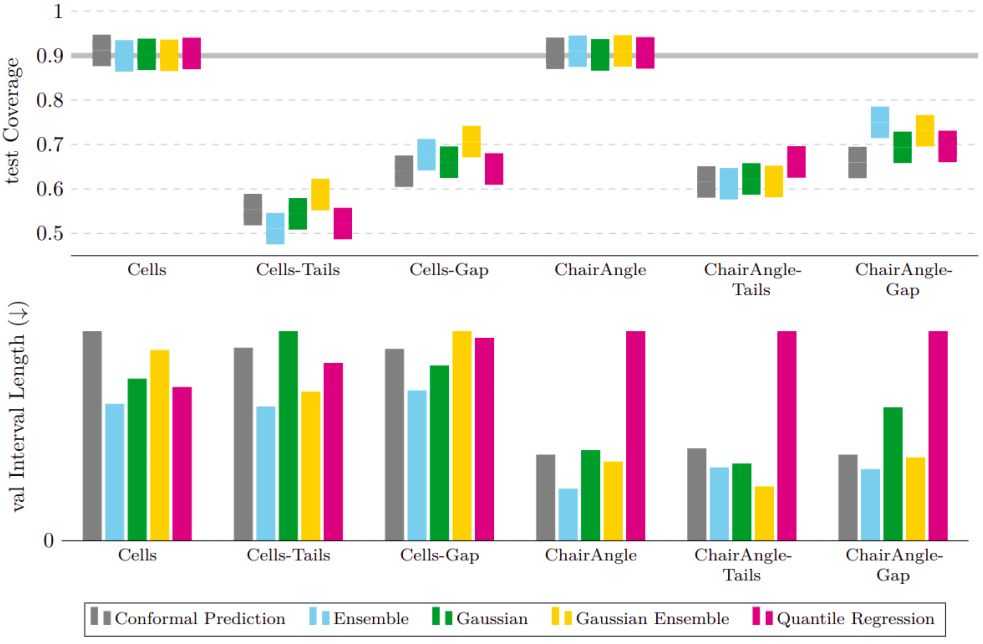
Gaussian.

Gaussian ensemble.

Quantile regression.

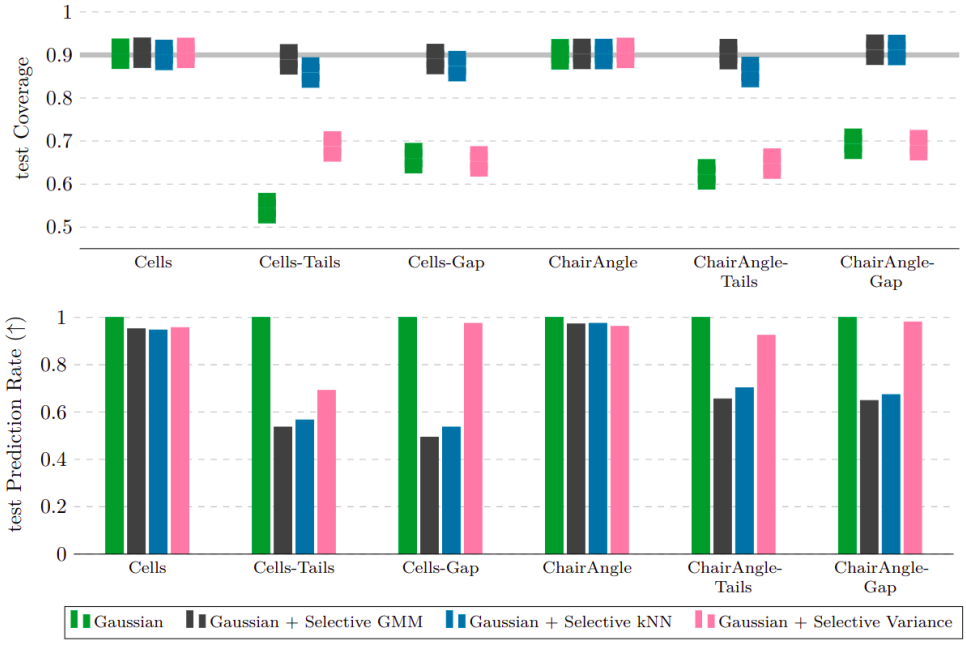
All methods output a prediction and a 90% *prediction interval for each input. We calibrate the intervals such that exactly* 90% *coverage is obtained on the val set.*

Results: Synthetic Datasets

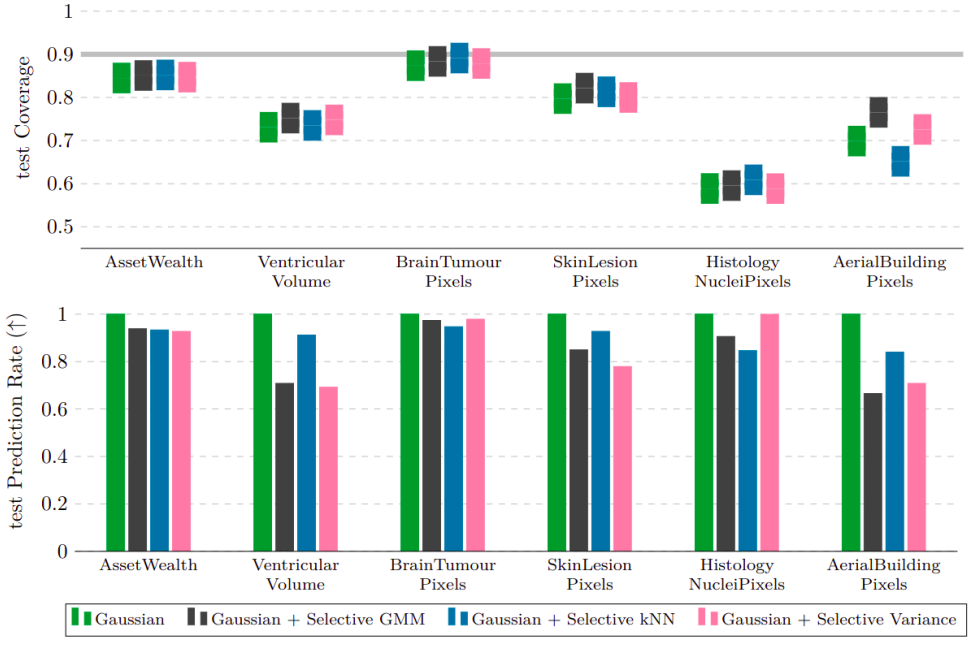


Results: Real-World Datasets

Results: Synthetic Datasets (Selective Prediction)



Results: Real-World Datasets (Selective Prediction)



Contact

Fredrik K. Gustafsson, Uppsala University
fredrik.gustafsson@it.uu.se
www.fregu856.com