# Computationally Efficient Classifiers with Frequentist Bounds on Prediction Errors

Shreeram Murali [1] Cristian R. Rojas [2] Dominik Baumann [1]

[1]**Aalto University** and the **Finnish Center of Artificial Intelligence (FCAI)**, Finland [2]**KTH Royal Institute of Technology**, Sweden

## Highlights

We propose a computationally efficient and effective classification algorithm based on the **Nadaraya-Watson Estimator** [1] that provides frequentist uncertainty bounds on its prediction estimates.

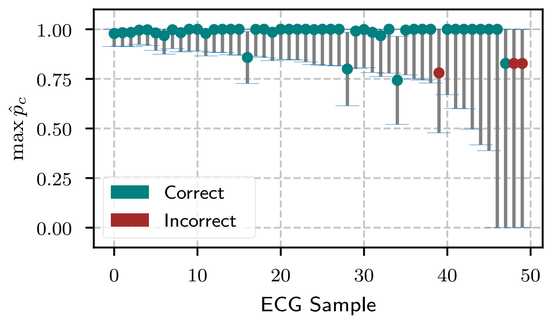



Figure 1. ECG predictions.

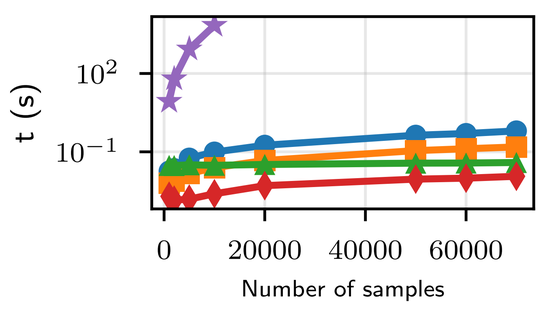



Figure 2. Total time vs training size.

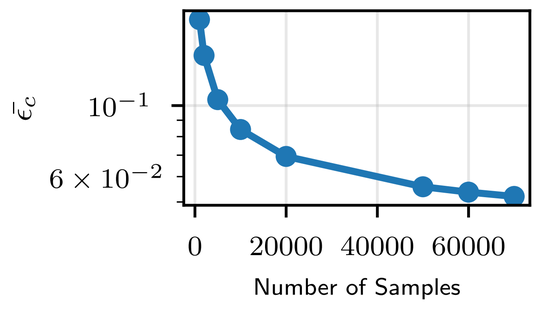



Figure 3. Bounds tightening with increased training size.

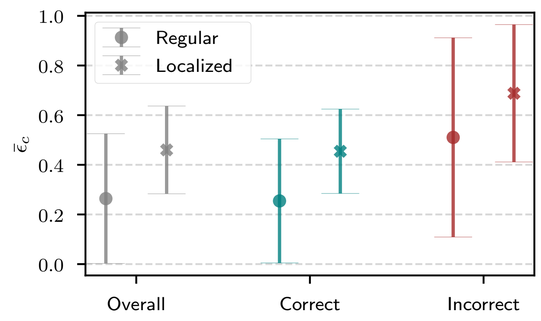



Figure 4. Mean bound for correct and incorrect predictions.

Regular NWC Dyadic NWC Localized NWC Logistic Regression CME

- Existing classifiers with frequentist bounds have poor computational scalability, such as Gaussian Process (GP) classification or **Conditional Mean Embeddings** (CMEs) [2].
- Unlike predecessor methods (CMEs, GPs) that scale with $\mathcal{O}(n^3)$, we develop a novel classifier that scales sublinearly.
- Proposed method jointly addresses *accuracy* ($> 96\%$), *computational efficiency* ($\mathcal{O}(n)$ to $\mathcal{O}(\log n)$), and provides theoretically guaranteed, actionable *bounds on prediction uncertainty*.

## Problem Definition

- **Task:** Given training data $\mathcal{D} = \{(y_i, c_i)\}_{i=1}^n$ with $y_i \in \mathbb{R}^d$, $c_i \in \mathbb{N}$, predict the class label $c$ for a new input $y$.
- **Objective:** Efficiently estimate $\hat{p}_c(y) \coloneqq \mathbb{P}(C = c \mid Y = y, \mathcal{D})$.
- **Requirements:** Provide frequentist confidence bounds for predictions; linear or sublinear scaling in $n$; support multi-class classification.

## Nadaraya-Watson Classifier

The Nadaraya-Watson Estimator (NWE) is a non-parametric kernel regression method used here to estimate class probabilities for classification.

- For a test input $y$, the estimated probability for class $c$ is:

$$\hat{\mathbf{p}}_c(y) = \frac{\sum_{i=1}^n K_\lambda(y, y_i)\mathbb{1}[c_i = c]}{\sum_{i=1}^n K_\lambda(y, y_i)}$$

- $K_\lambda(y, y_i)$: kernel function (e.g., Gaussian) that conveys similarity between $y$ and $y_i$
- $\mathbb{1}[c_i = c]$: indicator function for class match

## Uncertainty and Error Sources

- **Uncertainty sources:** *bias* due to model assumptions and kernel choice; *sampling error* due to finite training data.
- We bound the estimate as the sum of bias and sampling error.

$$\underbrace{|p_c(y) - \bar{p}_c(y)|}_{\text{bias}} + \underbrace{|\bar{p}_c(y) - \hat{p}_c(y)|}_{\text{sampling error}} \leq \epsilon_c(y, \delta, n) = \beta\lambda + 2\sigma\frac{\alpha_n(y, \delta)}{\kappa_n(y)}$$

- Here, $\bar{p}_c(y)$ is a virtual estimate we could determine if we had true probabilities instead of discrete labels

### Lemma 1: Bias for Lipschitz-continuous data distributions

*The bias of the Nadaraya-Watson classifier is bounded by*

$$|p_c(y) - \bar{p}_c(y)| \leq L\lambda$$

*where $L$ is the Lipschitz constant and $\lambda$ is the kernel bandwidth.*

### Lemma 2: Bias for distributions separated by a margin

*The bias of the Nadaraya-Watson classifier is bounded by*

$$|p_c(y) - \bar{p}_c(y)| \leq \frac{\lambda}{\gamma}$$

*where $\gamma$ is the smallest Euclidean distance between two samples $y$ and $y'$ with mismatched labels, and $\lambda$ is the kernel bandwidth.*

### Lemma 3: Data-dependent Statistical Error in Sampling

*With probability at least $1 - \delta$,*

$$|\bar{p}_c(y) - \hat{p}_c(y)| \leq \begin{cases} \frac{2\sigma}{\kappa_n(y)}\sqrt{\log\left(\frac{\sqrt{2}}{\delta}\right)} & \text{if } 0 < \kappa_n(y) \leq 1, \\ \frac{2\sigma}{\kappa_n(y)}\sqrt{k_n(y)\log(\delta^{-1}\sqrt{1+\kappa_n(y)})} & \text{if } \kappa_n(y) > 1. \end{cases}$$

*for $n$ samples.*

## Computational Efficiency Improvements

- We propose two computationally efficient variants: **Dyadic NWC**, with pre-computed hash-table construction, and **Localized NWC**, with *k*-nearest neighbours approach.
- Both methods reduce fit and prediction time from $\mathcal{O}(n^3)$ to sublinear complexity.

| Method | Fit | Prediction |
|---|---|---|
| Regular | – | $\mathcal{O}(n)$ |
| *Dyadic* | $\mathcal{O}(n)$ | $\mathcal{O}(\log n)$ |
| *Localized* | $\mathcal{O}(n \log n)$ | $\mathcal{O}(k + \log n)$ |

## Assumptions and Data Distributions

- We generate synthetic data to match our assumptions about the underlying distribution of the true probability $p_c$.
- We make regularity assumptions on the underlying true function $p_c(y)$, the nature of measurements, and the nature of the user-defined kernel $K_\lambda(y, y_i)$.

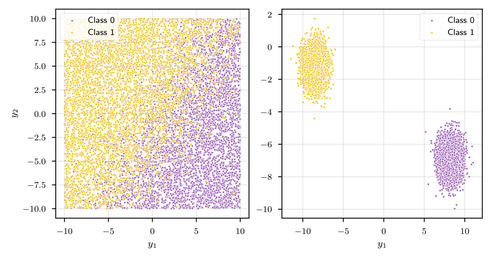


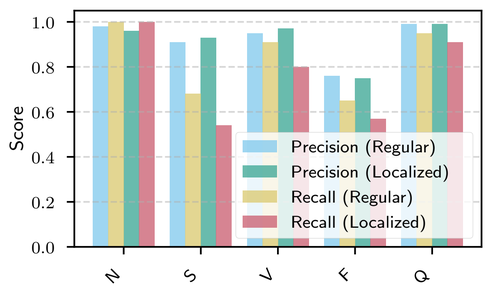

Figure 5. Overlapping and separable distributions.

We assume distributions are either **overlapping**, exhibiting Lipschitz-continuity with known constant $L$, or **separated** by a known margin $\gamma$.

## Arrhythmia Detection: ECG Results

- We evaluate the proposed classifier on a popular ECG database, the MIT-BIH dataset [3].
- Our classifier shows strong precision and recall scores, with higher uncertainty intervals associated with incorrect predictions.

| | **Accuracy** (%) | **Runtime** (s) |
|---|---|---|
| Regular | 96.2 | 16.91 |
| *Localized* | 97.8 | 2.01 |
| Logistic Reg. | 91.0 | 6.53 |
| CME | 92.0 | 8618.85 |



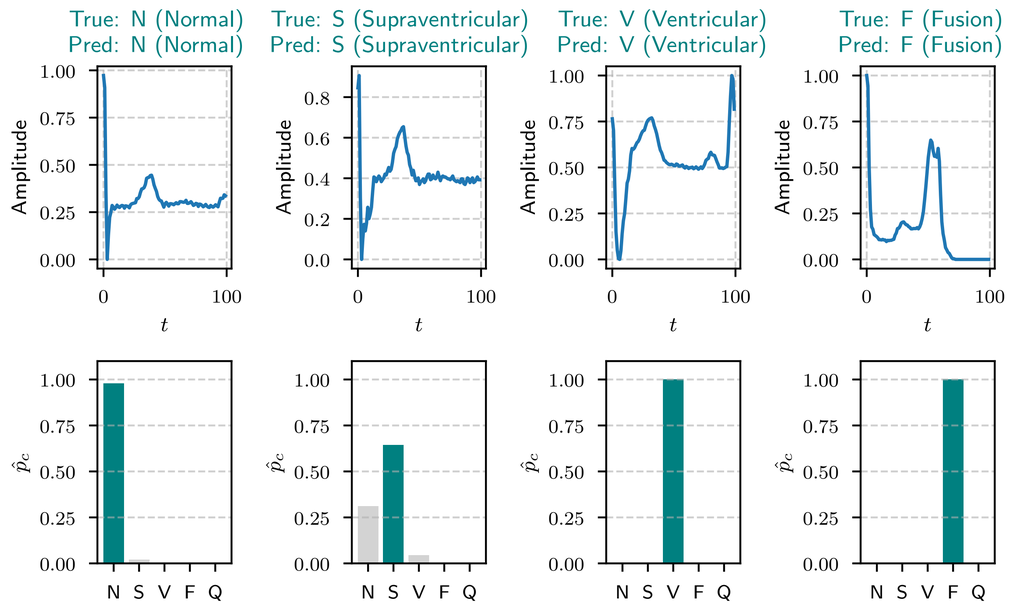




Figure 6. Top: precision and recall scores for the regular and localized classifiers across 5 heartbeat classes; bottom: selected ECG waveforms and their corresponding class probabilities.

## References

[1] D. Baumann, K. Kowalczyk, C. R. Rojas, K. Tiels, and P. Wachel. "Safety and optimality in learning-based control at low computational cost". In: *IEEE Transactions on Automatic Control* (2025).

[2] D. Baumann and T. B. Schön. "Safe reinforcement learning in uncertain contexts". In: *IEEE Transactions on Robotics* 40 (2024).

[3] G. Moody and R. Mark. "The impact of the MIT-BIH arrhythmia database". In: *IEEE Engineering in Medicine and Biology Magazine* 20.3 (2001).

Scan the QR code to read more!