Bootstrapping-Based Regularisation for Stable Deep Learning Clinical Prediction Models

Motivation & Contribution. Deep learning clinical prediction models have increasingly been applied for patient risk estimation, yet their outputs can be unstable when trained on different samples of the same population, thus undermining trust in clinical settings. Ensemble methods such as bagging can reduce this instability, but they require training and maintaining many models, which reduces interpretability, one of the key barriers to successful health AI implementation. We address this gap by introducing a *bootstrapping-based regularisation* that embeds stability directly into the training of deep neural networks (DNNs). By penalising divergence between predictions from the original training data and those from bootstrapped datasets, our method achieves ensemble-like robustness while remaining efficient and interpretable.

Methods & Data. Given training data $D = \{(x_i, y_i)\}_{i=1}^{N}$ with binary outcomes $y_i \in \{0, 1\}$ and a prediction model $f_\theta(x)$, we minimise the loss function

$$\mathcal{R}(\theta) = \mathcal{L}_\theta(D) + \frac{\lambda}{N} \sum_{i=1}^{N} \mathbb{E}_{B(D)} \Big[d\Big(f_{\hat{\theta}_b}(\mathbf{x}_i), f_\theta(\mathbf{x}_i) \Big) \Big]$$

where $\mathcal{L}_\theta(D)$ represents the standard binary cross-entropy loss and the second part is our additional regularisation. The stability penalty is defined as $d(f_{\hat{\theta}_b}(x), f_\theta(x)) = \| \log f_{\hat{\theta}_b}(x) - \log f_\theta(x) \|$, which measures how much the current model's predictions diverge from those produced by bootstrap-trained models $f_{\hat{\theta}_b}$ based on resamples $D^{(b)} \sim B(D)$.

Our *stable* model is a two-hidden-layer feed-forward DNN with a sigmoid activation, optimised using Adam. The expectation was approximated using predictions from 100 models drawn from a pool of 200 models trained on bootstrapped versions of the original datasets. We evaluated on three clinical datasets: *GUSTO-I* (n=40,830), *Framingham* (n=4,434), and *SUPPORT* (n=9,103). The *stable* model was compared to a *standard* DNN without the regularisation. Metrics assessed stability (mean absolute difference (MAD) between the *stable* model's predictions and median of the bootstrapped predictions, proportion of significantly *deviating predictions* from the median), discrimination (AUC), and feature attribution consistency between the *standard* model and *stable* model (SHAP correlations).

Results. Our *stable* model consistently reduced instability while maintaining discrimination and attribution consistency. Table 1 showcases some of the key results.

Dataset	**MAD** (↓)	**Deviating Pred. (%)** (↓)	**AUC** (↑)	**SHAP Corr.** (↑)
GUSTO-I	0.059 → 0.019	87.1 → 13.9	0.811 → 0.810	0.894
Framingham	0.088 → 0.057	55.0 → 21.4	0.810 → 0.815	0.965
SUPPORT	0.092 → 0.071	57.7 → 40.2	0.614 → 0.643	0.529

Table 1: Prediction stability improvements of the stable model relative to a standard DNN; SHAP correlation indicates per-participant attribution agreement. We report the change from a standard DNN to the stable model as Std → Stable. For each metric, arrows indicate the desired direction of better performance.

Conclusion. Embedding bootstraping-based regularisation as a training objective yields deep models with substantially improved individual-level prediction stability, whilst preserving AUC and strong SHAP concordance. By adjusting the regularisation strength λ, our framework spans a continuum between a standard model ($\lambda = 0$) and a bagging model ($\lambda \to \infty$). This allows users to tune the trade-off between predictive performance and stability while retaining the simplicity of a single, interpretable model. This stability-centric training objective addresses a key barrier of deep learning deployment in healthcare by producing risk estimates that are accurate, stable and reproducible.