Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Uncertainty-Quantified Gaussian Process Classification of Torsades de Pointes Risk from Pseudo-Coloured ECGs: External Validation in a UK Biobank Breast Cancer Cohort

Mohand Alhjaili
Newcastle University
Alaa Alahmadi
Alaa.Alahmadi@newcastle.ac.uk

Newcastle University

Research Article

Keywords: Torsades de Pointes, Electrocardiography, Gaussian process classification, Uncertainty quantification, External validation, Cardio-oncology

Posted Date: November 3rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11569821/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Machine-learning models for electrocardiogram (ECG)-based risk assessment are usually evaluated only on their development population, so their behaviour under population shift is rarely known. We examined this problem for Torsades de Pointes (TdP), a life-threatening arrhythmia associated with delayed ventricular repolarisation. A Bayesian Gaussian Process Classifier (GPC) with an Automatic Relevance Determination (ARD) Matérn (v = 2.5) kernel was trained to classify TdP risk, with labels defined by the QT nomogram, from 25 pseudo-coloured ECG features of healthy volunteers receiving QT-prolonging drugs (ECGRDVQ; 5,219 ECGs, 180 at risk of TdP) and applied without retraining to an independent UK Biobank cohort of patients with breast cancer (2,478 ECGs, 59 at risk of TdP). This population was chosen because it is commonly exposed to cardiotoxic, QT-prolonging therapies and is predominantly female, an established TdP risk factor. We also applied explainability analyses and decomposed predictive uncertainty. Internally, Balanced Accuracy was 0.987, sensitivity 0.978 and specificity 0.997. Externally, Balanced Accuracy was 0.835, sensitivity 0.814 and specificity 0.856. ARD length-scales, permutation importance and SHapley Additive exPlanations (SHAP) identified a single dominant feature, InflectionPoint, whose specificity declined externally: ECGs without TdP risk with InflectionPoint ≥ 4.5 rose from 7.3% to 15.1%. Aleatoric uncertainty exceeded epistemic uncertainty (0.167 versus 0.115 nats), indicating that errors arose mainly from class overlap rather than limited model knowledge. The model retained most of its sensitivity externally, and the analyses showed why specificity fell and what deployment requires: more discriminative features and population-representative training data.

1 Introduction

Torsades de Pointes (TdP) is a polymorphic ventricular tachycardia associated with QT-interval prolongation and delayed ventricular repolarisation caused by abnormal cardiac ion-channel function [1, 2]. Although episodes often terminate spontaneously, TdP can degenerate into ventricular fibrillation and sudden cardiac death, including in individuals without structural heart disease [1, 2]. The QT interval is the standard electrocardiogram (ECG) marker of delayed repolarisation, but its accurate measurement depends on locating the end of the T wave, which becomes unreliable when drug-induced changes such as T-wave flattening, notching or fusion with the U wave are present [2, 3]. At-Risk ECGs are also rare: in a placebo-controlled trial of four QT-prolonging drugs, only 180 of 5,050 recordings were classified as At-Risk [3]. Machine-learning models that combine ECG-derived and clinical features have been proposed, and in one large retrospective study such models outperformed established clinical risk scores for predicting drug-induced QT prolongation [4].

Alahmadi et al. proposed a pseudo-colouring approach that represents ventricular repolarisation relative to the QT nomogram on a nine-level colour scale [2, 3]. ECGs with a normal QT interval are dominated by cool pseudo-colours, whereas ECGs at risk of TdP contain a greater proportion of warm pseudo-colours [2]. Area-under-the-curve (AUC), derivative-based and InflectionPoint features derived from this representation describe repolarisation morphology more fully than the corrected QT interval (QTc) alone [3].

Predictive uncertainty is equally important in this setting. Bayesian models produce probabilistic predictions in which uncertainty is represented explicitly [5]. This uncertainty is commonly divided into epistemic uncertainty, which arises from limited model knowledge and can in principle be reduced with more data, and aleatoric uncertainty, which reflects inherent variability or class overlap in the data [6]. Reliable uncertainty quantification is increasingly recognised as a requirement for machine-assisted medical decision-making [7], not least because model performance can deteriorate substantially when deployment data differ from the training population [8]. Such dataset shift is a particular concern in cardio-oncology. Patients with breast cancer are a clinically important example: they are commonly treated with potentially cardiotoxic therapies, including anthracyclines, anti-HER2 agents and QT-prolonging drugs such as ribociclib, for which QTc monitoring is recommended [9, 10], and female sex is itself an established risk factor for drug-induced TdP [11]. For these reasons, a breast cancer cohort was chosen as the external validation population in this study.

Gaussian Process Classification (GPC) suits this problem for three reasons: it is a Bayesian, non-parametric method that produces probabilistic predictions with explicit uncertainty; it can be trained on the small, imbalanced datasets typical of TdP risk, whereas deep learning models generally require far larger training sets; and with Automatic Relevance Determination (ARD), its feature-specific length-scales indicate the relative relevance of each input [12]. However, ARD-enabled GPC has received little attention for ECG-based TdP risk prediction, and, more generally, models of drug-induced QT prolongation are rarely tested on an independent population that differs from the one used for development. When such a test does reveal a loss of performance, explainability and uncertainty analyses can show why, which is the information needed to decide whether, and how, a model can be deployed.

This study therefore develops an ARD Matérn GPC for binary TdP risk classification from 25 pseudo-coloured ECG features. Whereas our earlier algorithm applied expert-derived rules to these features [3], the present model learns their relationship with TdP risk from data, and the study makes four contributions. First, it develops and evaluates the model on ECGs from a drug trial in healthy volunteers and then tests the fixed model, without retraining, on an independent breast cancer cohort from the UK Biobank. Second, it uses ARD length-scales, permutation importance and Kernel SHapley Additive exPlanations (SHAP) to identify the features that drive predictions in each cohort. Third, it examines whether training on samples from both cohorts improves transfer. Fourth, it decomposes predictive uncertainty into aleatoric and epistemic components to establish whether errors under population shift reflect limited model knowledge or overlap between classes.

The remainder of the paper is organised as follows. Section 2 describes the datasets, model and validation design; Section 3 reports the results; Section 4 discusses their implications and limitations; and Section 5 concludes.

2 Methods

2.1 Study design

The study followed a three-stage design. First, a probabilistic classifier was developed and evaluated on an internal cohort using stratified cross-validation. Second, the fixed model was applied to an independent, unseen external cohort to assess generalisability. Third, samples from both cohorts were pooled to examine whether joint training improved performance across the two populations.

2.2 Datasets

The model was developed on the ECG Effects of Ranolazine, Dofetilide, Verapamil, and Quinidine in Healthy Subjects (ECGRDVQ) database, which is publicly available on PhysioNet [13, 14]. ECGRDVQ contains 12-lead ECGs from 22 healthy volunteers in a randomised, double-blind, five-period crossover trial (ClinicalTrials.gov NCT01873950) in which each participant received single doses of dofetilide, quinidine, ranolazine, verapamil and placebo; triplicate 10-second ECGs were extracted at 16 time points over each 24-hour period, giving 5,232 recordings [13]. After 13 recordings with missing QT values were excluded, 5,219 ECGs were included, of which 180 were At-Risk and 5,039 No-Risk (prevalence 3.45%). Following standard terminology for prediction-model validation, we refer to ECGRDVQ as the internal (development) cohort.

The external cohort comprised 2,478 ECGs from patients with breast cancer in the UK Biobank, of which 59 were At-Risk and 2,419 No-Risk (prevalence 2.38%); its selection is described in Section 2.3. This population was chosen because patients with breast cancer are commonly exposed to cardiotoxic and QT-prolonging therapies [9, 10] and are predominantly women, who are at higher risk of drug-induced TdP than men [11]. It differs from the internal cohort in several clinically important respects: it consists of older patients with cancer rather than young healthy volunteers, its ECGs were recorded during routine imaging assessments rather than after controlled drug exposure, and each participant contributes at most two recordings rather than hundreds. The external cohort was reserved entirely for external validation: it was not used for model development, hyperparameter selection or threshold optimisation. Its separate use in the re-training experiment is described in Section 2.6.3.

2.3 UK Biobank breast cancer cohort selection

UK Biobank is a prospective cohort of approximately 500,000 adults aged 40–69 years who were recruited at 22 assessment centres across the United Kingdom between 2006 and 2010, and whose records are linked to national hospital, cancer-registry and death-registry data [15]. Participants with breast cancer were identified from linked hospital inpatient records (Field 41270, diagnoses – ICD-10, which combines the main and secondary diagnoses in Fields 41202 and 41204) using the ICD-10 code C50 (malignant neoplasm of breast), including all subcodes C50.0–C50.9. This identified 13,024 participants (12,947 women and 77 men; mean age at recruitment 57.5 years, SD 7.5, range 40–70), of whom 12,704 had C50 recorded as a main diagnosis. Carcinoma in situ (D05) was not used as an inclusion code, and no participant was included on the basis of D05 alone; 1,256 of the included participants also had a D05 code recorded alongside C50. Men with breast cancer were not excluded.

Resting 12-lead ECGs were obtained from the UK Biobank imaging assessment visit (Field 20205, ECG at rest, 12-lead), where they were recorded at rest and stored as XML files containing the lead traces. ECGs

from both the initial imaging visit (instance 2, from 2014) and the first repeat imaging visit (instance 3, from 2019) were used, so a participant could contribute up to two recordings. Participants with a C50 diagnosis were linked to their ECG records using the UK Biobank participant identifier.

Of the 13,024 participants with a C50 diagnosis, those with at least one available resting 12-lead ECG were included, giving a final external cohort of 2,478 ECGs. Pseudo-coloured features were derived using the same procedure as for the internal cohort [3]. At-Risk labels were assigned with the same QT-nomogram criteria as for the internal cohort, by plotting the QT interval against heart rate, both taken from the UK Biobank ECG records, on the QT nomogram [16]. This yielded 59 At-Risk and 2,419 No-Risk ECGs.

2.4 Features

The features follow the pseudo-colouring approach developed by Alahmadi et al. [2, 3]. In this representation, the repolarisation (ST–T) segment of each beat is coloured according to its timing relative to the QT nomogram [16], a clinical tool that defines the QT interval at risk of TdP as a function of heart rate. For each beat, the at-risk QT value at the corresponding heart rate is read from the nomogram, and nine colour bands are anchored to it in 40 ms steps: purple, blue, green, lime and yellow begin 240, 200, 160, 120 and 80 ms before the at-risk value, orange 40 ms before it, dark orange at the at-risk value, and red and dark red 40 and 80 ms beyond it [2]. A normal QT interval therefore ends in the cool bands, whereas a prolonged, at-risk QT interval extends into the warm bands [2, 3]. Because the bands move with heart rate, the representation does not require a heart-rate-corrected QT interval, and because it uses the whole repolarisation segment rather than a single T-wave end point, it is less sensitive to abnormal T-wave morphology [3].

Each ECG is described by 25 features derived from this representation (Table A1, Appendix). First, the coloured T-wave (repolarisation) area of each beat is partitioned into the nine colour segments of the pseudo-colouring scale, and the area under the curve (AUC) of each segment is calculated and summed across all beats in the recording. This gives nine ∑AUC features, one for each colour (purple, blue, green, lime, yellow, orange, dark orange, red and dark red). Plotted in colour order, they form a global representation graph of the recording that shows how much of the T-wave area falls into each colour band: in ECGs at risk of TdP, the area shifts from the cool to the warm bands [3].

Second, the first derivative of this graph is taken between each pair of adjacent colours, $\sum AUC_{k+1} - \sum AUC_k$ (k = 1, …, 8, from purple to dark red), giving eight features named after the two colours involved; for example, Blue_Purple = $\sum AUC_{blue} - \sum AUC_{purple}$. These describe how the area changes from one colour to the next. Third, the second derivative is taken across each set of three adjacent colours, $\sum AUC_{k+1} - 2\sum AUC_k + \sum AUC_{k-1}$, giving seven features named after the three colours; for example, Purple_Blue_Green = $\sum AUC_{green} - 2\sum AUC_{blue} + \sum AUC_{purple}$. These describe the curvature of the graph: a positive Purple_Blue_Green value means that the graph bends upwards at the blue band, and a negative value that it bends downwards. Finally, the InflectionPoint index, developed in our previous work [3], gives the

position on the colour scale (1 = purple to 9 = dark red) at which the curvature of the graph changes sign. Figures 1 and 2 show the pseudo-coloured ECG and its global representation graph for ECGs with a normal QT interval and with a QT interval at risk of TdP, respectively; reference [3] describes the InflectionPoint index and the pseudo-coloured ECG in full.

In our previous work [3], these features were interpreted by a rule-based algorithm that applied expert-derived decision rules to the AUC and derivative features. The present study instead uses probabilistic machine learning to learn the relationship between the 25 features and TdP risk directly from the data. This yields predicted probabilities with quantified uncertainty and a data-driven measure of the relevance of each feature (Section 2.5).

The features were pre-computed and supplied with the datasets rather than extracted from raw ECG signals in this study. All features were standardised to zero mean and unit variance using parameters estimated on the training data only; these parameters were then applied to the corresponding validation or test data to prevent information leakage.

2.5 Model

A GPC was used to model the relationship between the 25 features and binary TdP risk. GPC places a Gaussian process (GP) prior over a latent function, which is mapped to class probabilities through a logistic link function [12]. Because the posterior is not analytically tractable for classification, inference was performed with the Laplace approximation [12]. The covariance function combined a Matérn kernel ($\nu = 2.5$) with ARD, which assigns a separate length-scale to each feature [12], and a White kernel that models an additional noise component and improves numerical stability [17]. Kernel hyperparameters were optimised by maximising the Laplace-approximated log marginal likelihood [12, 17]. Shorter length-scales indicate that the latent function is more sensitive to the corresponding feature. The Matérn kernel was chosen because it imposes weaker smoothness assumptions than the squared-exponential kernel: with $\nu = 2.5$, sample functions are twice differentiable rather than infinitely smooth [12].

To address class imbalance and the cubic computational cost of Gaussian process inference, all At-Risk samples were retained during training together with a random subsample of No-Risk cases, and additional weight was given to the minority class. The model was implemented in Python using scikit-learn [17], and random seeds were fixed to support reproducibility.

The enhanced configuration (25 features, ARD Matérn kernel) was compared with a previous configuration that used 24 features (without InflectionPoint) and an isotropic Matérn ($\nu = 2.5$) kernel, at the same decision threshold of 0.55.

2.6 Validation design

Because accuracy is uninformative under severe class imbalance, performance was assessed primarily with Sensitivity, Specificity and Balanced Accuracy (their mean). Unlike the positive predictive value (PPV), these measures do not depend on the prevalence of the At-Risk class, which was low in both

cohorts and differed between them; PPV was therefore not used as a performance metric, and the practical burden of false positives is instead reported directly as counts. The F2 score was used as the main criterion for selecting the decision threshold. It belongs to the Fβ family, $F\beta = (1 + \beta^2) \cdot PPV \cdot Sensitivity / (\beta^2 \cdot PPV + Sensitivity)$, in which β sets the relative importance of Sensitivity; with β = 2, Sensitivity is treated as twice as important as PPV [19]. This choice reflects the asymmetric clinical costs of the two types of error in TdP risk assessment. A missed At-Risk ECG (false negative) may allow a QT-prolonging drug to be continued in a patient at risk of a potentially fatal arrhythmia, whereas a false alarm (false positive) typically leads to a repeat ECG or expert review of the QT interval. F2 therefore favours Sensitivity while, unlike Sensitivity alone, still penalising thresholds that generate excessive false positives; trade-offs of this kind are more informative than receiver operating characteristic (ROC)-based measures when the positive class is rare [18]. For the external validation, 95% confidence intervals (CIs) for Sensitivity, Specificity and negative predictive value (NPV) were computed from the confusion matrix using the Wilson score method.

2.6.1 Internal validation

The internal cohort was evaluated using stratified five-fold cross-validation, which preserves the proportion of At-Risk and No-Risk cases in each fold. The decision threshold was selected by evaluating a range of thresholds on the out-of-fold posterior probability of At-Risk status and choosing the value that maximised F2. The resulting configuration and threshold (0.55) were then fixed.

2.6.2 External validation

A final model was trained on the internal cohort, using the subsampling scheme described in Section 2.5, and applied to the external cohort without refitting or re-tuning. Feature standardisation used parameters estimated on the internal training data only, and the 0.55 threshold was retained, so that no external labels were used at any stage of model optimisation.

2.6.3 Re-training experiment

To examine whether exposure to both populations improves performance, a pooled set of 600 samples was constructed, comprising 300 samples from each cohort and including all 239 At-Risk cases (180 internal and 59 external). Stratified five-fold cross-validation was performed on the pooled data and repeated across multiple random seeds, and out-of-fold predictions were reported separately for each cohort. Because all At-Risk cases were retained, At-Risk prevalence in the pooled subsets (60.0% internal and 19.7% external) was much higher than in the original cohorts, so prevalence-dependent measures in this experiment are not directly comparable with those in the external validation. As this experiment uses external samples for training, it is not an external validation and is reported as a separate analysis.

2.7 Uncertainty and explainability

Predictive uncertainty was decomposed using Monte Carlo samples drawn from the approximate (Laplace) posterior over the latent function. Total predictive entropy was separated into aleatoric

uncertainty, the expected entropy of the predictive distribution, and epistemic uncertainty, the mutual information between the prediction and the latent function, following Houlsby et al. [20] and Depeweg et al. [21]. All values are reported in nats. Feature relevance was assessed using the fitted ARD length-scales [12], permutation importance [22] and Kernel SHAP [23]; SHAP values were also computed separately for the internal and external samples of the pooled model.

2.8 Exploratory feature-space analyses

Feature distributions were explored using principal component analysis (PCA), uniform manifold approximation and projection (UMAP) and a further two-dimensional embedding, together with Pearson correlation matrices of the 25 features in each cohort. Pearson correlations between each feature and the out-of-fold predictive entropy of the internal model were also computed. As a diagnostic only, the cohorts were re-embedded after standardising each cohort with its own mean and variance; this analysis was not used for model development or evaluation.

3 Results

3.1 Cohorts

Both cohorts were highly imbalanced, with At-Risk ECGs forming a small minority (Table 1).

Table 1
Cohort composition and class distribution

Cohort	Total	At-Risk	No-Risk	Prevalence
Internal ECGRDVQ cohort (healthy volunteers)	5,219	180	5,039	3.45%
External breast cancer ECG cohort (UK Biobank)	2,478	59	2,419	2.38%

3.2 Internal performance

Discrimination was strong and consistent across folds, with fold-level Balanced Accuracy of 0.97–0.99, Sensitivity of 0.94–1.00 and Specificity of 0.99 (Table 2). The 0.55 threshold achieved the highest F2 score in the threshold sweep. Sensitivity and Specificity remained high over a wide range of thresholds, and above approximately 0.65 Sensitivity and F2 declined as more At-Risk cases were missed. At the selected threshold, the model detected 97.8% of At-Risk cases with a Specificity of 0.997, corresponding to approximately 15 false positives among 5,039 No-Risk ECGs (Table 3).

Table 2
Internal five-fold cross-validation results (ARD Matérn v = 2.5 + White kernel, 25 features, threshold = 0.55)

Fold	Balanced Accuracy	Sensitivity	Specificity
1	0.97	0.94	0.99
2	0.98	0.97	0.99
3	0.98	0.97	0.99
4	0.99	1.00	0.99
5	0.99	1.00	0.99

Table 3
Overall internal cross-validation performance and threshold results (threshold = 0.55)

Metric	Overall
Balanced Accuracy	0.987
Sensitivity	0.978
Specificity	0.997
F2	0.966

The out-of-fold predicted probabilities showed the same pattern: most No-Risk cases received low probabilities, below the 0.55 threshold, whereas most At-Risk cases clustered at the high-probability end, and only a few samples lay close to the decision boundary. This is consistent with the high Sensitivity (0.978) and Specificity (0.997) in Table 3 and shows that the model separated the classes well despite an At-Risk prevalence of only 3.45%.

3.3 Effect of the enhanced ARD configuration

Compared with the previous 24-feature isotropic configuration at the same threshold (Table 4), the enhanced 25-feature ARD configuration improved both Sensitivity (+ 0.017) and Specificity (+ 0.010). Although small in absolute terms, the gain in Specificity reduced the false-positive rate from 1.3% to 0.3%, which matters when At-Risk ECGs are rare. Because InflectionPoint and the ARD kernel were introduced together, their individual contributions cannot be separated from this comparison.

Table 4
Comparison between the previous 24-feature isotropic configuration and the enhanced 25-feature ARD configuration

Metric	Before: 24-feature isotropic	After: 25-feature ARD	Change
Sensitivity	0.961	0.978	+ 0.017
Specificity	0.987	0.997	+ 0.010

3.4 External validation

When the fixed model was applied without retraining to the external cohort, performance decreased (Tables 5–7). Sensitivity fell to 0.814 (95% CI 0.696–0.893) and Specificity to 0.856 (0.841–0.869), indicating that the model retained a meaningful ability to distinguish At-Risk from No-Risk ECGs; of the 59 At-Risk cases, 48 were detected and 11 were missed. Because At-Risk ECGs are rare, however, the fall in Specificity had a large practical effect: 349 of the 2,419 No-Risk ECGs were classified as At-Risk, so false positives outnumbered true positives by more than seven to one. Conversely, only 11 of the 2,081 ECGs predicted to be No-Risk were At-Risk (NPV 0.995, 95% CI 0.991–0.997).

Table 5
External validation results on the breast cancer ECG cohort (threshold = 0.55; 95% Wilson confidence intervals in parentheses)

Metric	External result
Balanced Accuracy	0.835
Sensitivity	0.814 (0.696–0.893)
Specificity	0.856 (0.841–0.869)
NPV	0.995 (0.991–0.997)

Table 6
External validation confusion matrix (n = 2,478)

Actual No-Risk	Predicted No-Risk	Predicted At-Risk
	2,070	**349**
Actual At-Risk	11	48

Table 7
Internal versus external performance

Metric	Internal	External	Difference
Balanced Accuracy	0.987	0.835	−0.152
Sensitivity	0.978	0.814	−0.164
Specificity	0.997	0.856	−0.141

3.5 InflectionPoint dependence

External predictions depended strongly on the InflectionPoint feature: every externally predicted At-Risk case had InflectionPoint ≥ 4.5, whereas every missed At-Risk case fell below this value. The proportion of No-Risk cases with InflectionPoint ≥ 4.5 increased from 7.3% in the internal cohort to 15.1% in the external cohort, so the feature was less specific to the At-Risk class in the external population. Projections of the pooled feature space separated the two cohorts more clearly than the two classes (Fig. 3), consistent with a cohort-dependent shift in the feature distributions.

In the internal cohort, features within each colour-based group were strongly correlated, particularly between neighbouring waveform segments, whereas InflectionPoint correlated only weakly with most other features, indicating that it carries largely independent information (Fig. 4). The external cohort showed a broadly similar correlation structure (Fig. 5); the purple feature had zero variance in this cohort, so its correlations were undefined.

A two-dimensional embedding of the standardised feature vectors showed a clear distributional difference between the cohorts (Fig. 6). With standardisation based on the internal training data, as used in the external validation, the two cohorts occupied largely separate regions of the feature space. When each cohort was standardised with its own mean and variance, the overlap increased substantially, indicating that part of the cohort difference reflects shifts in feature location and scale. This diagnostic does not, however, resolve the underlying difference: per-cohort standardisation requires prior access to the target population, and some cohort-specific differences, including the behaviour of InflectionPoint, persisted.

3.6 Feature importance

In the ARD analysis, InflectionPoint had the shortest length-scale and was therefore the most relevant of the 25 features. Permutation importance and Kernel SHAP analysis of the pooled model (Fig. 7) also ranked InflectionPoint first. In the cohort-stratified SHAP analysis, InflectionPoint remained the most influential feature in both the internal and external samples, indicating that its dominance was not an artefact of training on the internal cohort alone.

InflectionPoint also had the strongest correlation with out-of-fold predictive entropy in the internal cohort ($r = -0.25$): higher InflectionPoint values were generally associated with lower entropy and therefore

greater model confidence. Several AUC and derivative features, including blue, Blue_Purple and Green_Lime_Yellow, showed weak to moderate correlations (|r| ≈ 0.15–0.21) (Fig. 8). These results are consistent with the ARD ranking and with the model's strong dependence on InflectionPoint in the external validation.

3.7 Re-training experiment

Pooled training achieved high fold-level performance, with Sensitivity of 0.83–0.96 and Specificity of 0.90–0.99 (Table 8). When results were stratified by cohort, performance remained higher on the internal samples (Table 9). On the external samples, Specificity increased from 0.856 with the fixed model to 0.913, whereas Sensitivity decreased from 0.814 to 0.712. At the original external prevalence, this gain in Specificity would reduce false positives from 349 to approximately 210 of the 2,419 No-Risk ECGs. Exposure to external samples therefore shifted the model towards fewer false positives at the cost of more missed At-Risk cases, without removing the difference between cohorts.

Table 8
Combined-cohort five-fold cross-validation results

Fold	Sensitivity	Specificity
1	0.938	0.958
2	0.833	0.944
3	0.958	0.903
4	0.917	0.903
5	0.936	0.986

Table 9
Combined-cohort out-of-fold performance, reported separately by cohort

Cohort	Sensitivity	Specificity
Internal (180/300 At-Risk)	0.983	0.992
External (59/300 At-Risk)	0.712	0.913

3.8 Uncertainty decomposition

On the pooled 600-sample dataset, aleatoric uncertainty (0.167 nats) exceeded epistemic uncertainty (0.115 nats), giving an aleatoric-to-epistemic ratio of approximately 1.46 (Table 10). Projection of the samples onto the first two principal components, coloured by predicted probability and by predictive entropy (Fig. 9), showed that most internal samples lay in a low-risk, low-uncertainty region, with uncertain cases concentrated near the boundary between No-Risk and At-Risk predictions. In the

external cohort, the transition between low- and high-risk predictions was broader and the high-entropy region larger.

Table 10
Aleatoric/epistemic uncertainty decomposition on the combined 600-sample dataset

Uncertainty component	Value
Epistemic uncertainty (EU)	0.115 nats
Aleatoric uncertainty (AU)	0.167 nats
AU/EU ratio	1.46×

External samples extended towards higher aleatoric uncertainty at relatively low epistemic uncertainty, whereas internal samples were concentrated at lower aleatoric values (Fig. 10). Thus, when the model was uncertain, this was mainly because the ECG features did not cleanly separate At-Risk from No-Risk cases, rather than because the model lacked knowledge.

4 Discussion

4.1 Findings

The central finding of this study is the marked gap between internal and external performance. The ARD Matérn GPC performed strongly on the internal cohort, but when the fixed model was applied without retraining to the UK Biobank breast cancer cohort, Specificity fell from 0.997 to 0.856 and Sensitivity from 0.978 to 0.814 (Table 7). Discriminative ability was therefore largely retained, but because At-Risk ECGs are rare (prevalence 2.38%), the fall in Specificity produced 349 false positives against 48 true positives, more than seven false alarms for every At-Risk ECG detected. The feature–outcome relationship learned internally was therefore only partly preserved in the external population. This is unsurprising given how different the two populations are: the model was developed on young healthy volunteers whose repolarisation was altered by controlled exposure to QT-prolonging drugs, and was applied to older patients with breast cancer, in whom age, comorbidity and cancer therapies all influence repolarisation morphology.

The explainability analyses point to a specific mechanism. The model depended heavily on InflectionPoint: every externally predicted At-Risk case had InflectionPoint ≥ 4.5, and every missed At-Risk case fell below this value. Consequently, an ECG dominated by cool colours could still be classified as At-Risk if its InflectionPoint fell within the learned decision region, whereas an ECG with larger warm-coloured AUC regions could be classified as No-Risk if its InflectionPoint remained below it. Because the proportion of No-Risk ECGs with InflectionPoint ≥ 4.5 doubled from 7.3% to 15.1%, the feature became less specific to the At-Risk class externally, which largely explains the excess of false positives.

InflectionPoint was also the most influential feature in both cohorts in the SHAP analysis of the pooled model (Fig. 7), so its dominance was not an artefact of internal-only training.

The uncertainty decomposition adds a further perspective. Aleatoric uncertainty exceeded epistemic uncertainty (0.167 versus 0.115 nats), suggesting that the external errors arose mainly from overlap between classes in the feature space rather than from model ignorance. Posterior confidence therefore does not guarantee correctness when the feature–outcome relationship shifts, and richer or more discriminative features, rather than more data from the same feature space alone, are likely to be needed. Consistent with this, training on both cohorts reduced false positives at the cost of some missed At-Risk cases (Table 9) but did not remove the difference between cohorts.

From a healthcare-informatics perspective, these results suggest that the model in its current form is better suited to ruling out risk than to confirming it. In the external cohort, only 11 of 2,081 ECGs predicted to be No-Risk were At-Risk (NPV 0.995), although this partly reflects the low prevalence, whereas a positive prediction would require confirmation, for example by expert review of the QT interval. Before deployment in a new population, local recalibration of the decision threshold, monitoring for shifts in key features such as InflectionPoint, and training data that represent the target population are likely to be necessary [8].

4.2 Strengths and limitations

The main strengths of this study are the use of an entirely unseen and clinically distinct external cohort, the strict separation of external data from model development and threshold selection, and the combination of performance evaluation with feature-relevance and uncertainty analyses that help to explain the model's behaviour.

Several limitations should be noted. Generalisation was assessed in only one external cohort, and the number of At-Risk cases was small (180 internal and 59 external); estimates of minority-class metrics are therefore imprecise, as the wide CI for external Sensitivity shows, and CIs were not estimated for the internal cross-validation results. ECGRDVQ contains repeated recordings from only 22 participants; when recordings from the same participant appear in both training and test folds, internal performance can be overestimated, which may contribute to the gap between internal and external results. The At-Risk label is derived from QT-based criteria rather than observed TdP events. The subsampling of No-Risk cases required for Gaussian process inference may affect the stability of the estimates, and the decision threshold was selected on the same out-of-fold predictions used to report internal performance, which may make the internal estimates slightly optimistic. In the re-training experiment, the pooled subsets had a much higher At-Risk prevalence than the original cohorts, so prevalence-dependent measures there are not directly comparable with those of the external validation. UK Biobank participants are, on average, healthier than the general UK population and predominantly of White European ancestry, which may limit generalisability. Breast cancer was identified from hospital inpatient records without linkage to cancer-registry data, so cases managed entirely outside hospital admission may have been missed. Finally, no prospective clinical validation was performed.

5 Conclusions

A Bayesian GPC with ARD trained on pseudo-coloured ECG features from a drug trial in healthy volunteers achieved strong internal performance for TdP risk classification and retained most of its sensitivity when applied, without retraining, to an independent UK Biobank breast cancer cohort, although Specificity fell and false positives increased substantially. Feature-relevance analysis traced this loss of Specificity to a single dominant feature, InflectionPoint, whose relationship with the At-Risk class shifted between populations, and the uncertainty decomposition showed that the remaining uncertainty reflected class overlap rather than a shortage of model knowledge. For healthcare informatics, the study shows that combining external validation with explainability and uncertainty decomposition does more than measure a drop in performance: it identifies the cause and indicates whether more data or better features are needed. Future work will evaluate richer repolarisation features, participant-level validation splits and further external cohorts, including patients receiving specific QT-prolonging cancer therapies.

Declarations

Competing interests

The authors have no relevant financial or non-financial interests to disclose.

Ethics approval

This study is a secondary analysis of existing, de-identified data. UK Biobank has approval from the North West Multi-centre Research Ethics Committee as a Research Tissue Bank (REC reference 21/NW/0157). The ECGRDVQ data come from a clinical trial (NCT01873950) conducted with ethics approval and written informed consent, as reported by Johannesen et al. [13], and are distributed in de-identified form by PhysioNet under the Open Data Commons Attribution License v1.0. The study was conducted in accordance with the principles of the Declaration of Helsinki; no additional ethical approval was required for the present analysis.

Consent to participate

All UK Biobank participants provided written informed consent. Participants in the ECGRDVQ trial provided written informed consent in the original study [13].

Consent for publication

Not applicable.

Funding

No specific funding was received for this work beyond the support for UK Biobank data access described in the Acknowledgements.

Author Contribution

A.A. supervised the study, provided the data and the code used to generate the pseudo-coloured ECG features, and edited and finalised the manuscript. M.A. applied the probabilistic machine learning models, performed the analyses, produced the results, prepared Figures 3–10 and wrote the first draft of the manuscript. All authors reviewed and approved the final manuscript.

Acknowledgements

This research was conducted using the UK Biobank Resource under Application Number 179597. Access to UK Biobank data was funded through The University of Manchester, with Newcastle University as a project partner.

Data Availability

The data are held by third parties and are not publicly available. UK Biobank data are available to bona fide researchers on application (https://www.ukbiobank.ac.uk). The ECGRDVQ database is freely available from PhysioNet (https://doi.org/10.13026/C2HP45). The pre-computed pseudo-coloured ECG features are available from the corresponding author on reasonable request.

Code availability

The analysis code is available from the corresponding author on reasonable request.

References

1. Yap YG, Camm AJ (2003) Drug induced QT prolongation and torsades de pointes. Heart 89:1363–1372. https://doi.org/10.1136/heart.89.11.1363
2. Alahmadi A, Davies A, Vigo M, Jay C (2020) Pseudo-colouring an ECG enables lay people to detect QT-interval prolongation regardless of heart rate. PLoS ONE 15:e0237854. https://doi.org/10.1371/journal.pone.0237854
3. Alahmadi A, Davies A, Royle J, Goodwin L, Cresswell K, Arain Z, Vigo M, Jay C (2021) An explainable algorithm for detecting drug-induced QT-prolongation at risk of torsades de pointes (TdP) regardless of heart rate and T-wave morphology. Comput Biol Med 131:104281. https://doi.org/10.1016/j.compbiomed.2021.104281

4. Jing L, Morland TB, Kelsey CR, Raghunath S, Pfeifer JM, Ruhl JA, Steinhubl S, Monfette M, Fornwalt BK, Haggerty CM (2026) Machine learning–enabled assessment of risk of drug-induced QT prolongation at the time of prescribing. Heart Rhythm 23:925–935. https://doi.org/10.1016/j.hrthm.2025.10.045
5. Chen IY, Joshi S, Ghassemi M, Ranganath R (2021) Probabilistic machine learning for healthcare. Annu Rev Biomed Data Sci 4:393–415
6. Der Kiureghian A, Ditlevsen O (2009) Aleatory or epistemic? Does it matter? Struct. Saf 31:105–112
7. Begoli E, Bhattacharya T, Kusnezov D (2019) The need for uncertainty quantification in machine-assisted medical decision making. Nat Mach Intell 1:20–23
8. Finlayson SG, Subbaswamy A, Singh K, Bowers J, Kupke A, Zittrain J, Kohane IS, Saria S (2021) The clinician and dataset shift in artificial intelligence. N Engl J Med 385:283–286
9. Lyon AR, López-Fernández T, Couch LS et al (2022) 2022 ESC Guidelines on cardio-oncology developed in collaboration with the European Hematology Association (EHA), the European Society for Therapeutic Radiology and Oncology (ESTRO) and the International Cardio-Oncology Society (IC-OS). Eur Heart J 43:4229–4361. https://doi.org/10.1093/eurheartj/ehac244
10. Porta-Sánchez A, Gilbert C, Spears D, Amir E, Chan J, Nanthakumar K, Thavendiranathan P (2017) Incidence, diagnosis, and management of QT prolongation induced by cancer therapies: a systematic review. J Am Heart Assoc 6:e007724. https://doi.org/10.1161/JAHA.117.007724
11. Makkar RR, Fromm BS, Steinman RT, Meissner MD, Lehmann MH (1993) Female gender as a risk factor for torsades de pointes associated with cardiovascular drugs. JAMA 270:2590–2597. https://doi.org/10.1001/jama.270.21.2590
12. Rasmussen CE, Williams CKI (2006) Gaussian Processes for Machine Learning. MIT Press, Cambridge, MA
13. Johannesen L, Vicente J, Mason JW, Sanabria C, Waite-Labott K, Hong M, Guo P, Lin J, Sørensen JS, Galeotti L, Florian J, Ugander M, Stockbridge N, Strauss DG (2014) Differentiating drug-induced multichannel block on the electrocardiogram: randomized study of dofetilide, quinidine, ranolazine, and verapamil. Clin Pharmacol Ther 96:549–558. https://doi.org/10.1038/clpt.2014.155
14. Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PC, Mark RG, Mietus JE, Moody GB, Peng C-K, Stanley HE (2000) PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation 101:e215–e220. https://doi.org/10.1161/01.CIR.101.23.e215
15. Sudlow C, Gallacher J, Allen N et al (2015) UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med 12:e1001779. https://doi.org/10.1371/journal.pmed.1001779
16. Chan A, Isbister GK, Kirkpatrick CMJ, Dufful SB (2007) Drug-induced QT prolongation and torsades de pointes: evaluation of a QT nomogram. QJM 100:609–615
17. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E (2011)

Scikit-learn: machine learning in Python. J Mach Learn Res 12:2825–2830

18. Davis J, Goadrich M (2006) The relationship between precision-recall and ROC curves. In: Proceedings of the 23rd International Conference on Machine Learning, pp. 233–240. ACM, New York
19. van Rijsbergen CJ (1979) Information Retrieval, 2nd edn. Butterworths, London
20. Houlsby N, Huszár F, Ghahramani Z, Lengyel M (2011) Bayesian active learning for classification and preference learning. https://doi.org/10.48550/arXiv.1112.5745. arXiv
21. Depeweg S, Hernández-Lobato JM, Doshi-Velez F, Udluft S (2018) Decomposition of uncertainty in Bayesian deep learning for efficient and risk-sensitive learning. Proc. Mach. Learn. Res. 80, 1184–1193 https://proceedings.mlr.press/v80/depeweg18a.html
22. Breiman L (2001) Random forests. Mach Learn 45:5–32
23. Lundberg SM, Lee S-I (2017) A unified approach to interpreting model predictions. Adv Neural Inf Process Syst 30:4765–4774

Appendix

Table A1
Definitions of the 25 pseudo-coloured ECG features. Feature names are those used in the figures; ∑AUC(colour) denotes the summed area under the curve of that colour segment

Group	Feature	Definition
Colour-band area (9)	purple	∑AUC of the purple segment, summed over all beats
	blue	∑AUC of the blue segment, summed over all beats
	green	∑AUC of the green segment, summed over all beats
	lime	∑AUC of the lime segment, summed over all beats
	yellow	∑AUC of the yellow segment, summed over all beats
	orange	∑AUC of the orange segment, summed over all beats
	dark_orange	∑AUC of the dark orange segment, summed over all beats
	red	∑AUC of the red segment, summed over all beats
	dark_red	∑AUC of the dark red segment, summed over all beats
First derivative (8)	Blue_Purple	∑AUC(blue) − ∑AUC(purple)
	Green_Blue	∑AUC(green) − ∑AUC(blue)
	Lime_Green	∑AUC(lime) − ∑AUC(green)
	Yellow_Lime	∑AUC(yellow) − ∑AUC(lime)
	Orange_Yellow	∑AUC(orange) − ∑AUC(yellow)
	DarkOrange_Orange	∑AUC(dark orange) − ∑AUC(orange)
	Red_DarkOrange	∑AUC(red) − ∑AUC(dark orange)
	DarkRed_Red	∑AUC(dark red) − ∑AUC(red)
Second derivative (7)	Purple_Blue_Green	∑AUC(green) − 2∑AUC(blue) + ∑AUC(purple)
	Blue_Green_Lime	∑AUC(lime) − 2∑AUC(green) + ∑AUC(blue)
	Green_Lime_Yellow	∑AUC(yellow) − 2∑AUC(lime) + ∑AUC(green)
	Lime_Yellow_Orange	∑AUC(orange) − 2∑AUC(yellow) + ∑AUC(lime)
	Yellow_Orange_DarkOrange	∑AUC(dark orange) − 2∑AUC(orange) + ∑AUC(yellow)

Group	Feature	Definition
	Orange_DarkOrange_Red	∑AUC(red) − 2∑AUC(dark orange) + ∑AUC(orange)
	DarkOrange_Red_DarkRed	∑AUC(dark red) − 2∑AUC(red) + ∑AUC(dark orange)
Inflection (1)	InflectionPoint	Position on the colour scale (1–9) where the curvature of the ∑AUC graph changes sign [3]

Figures

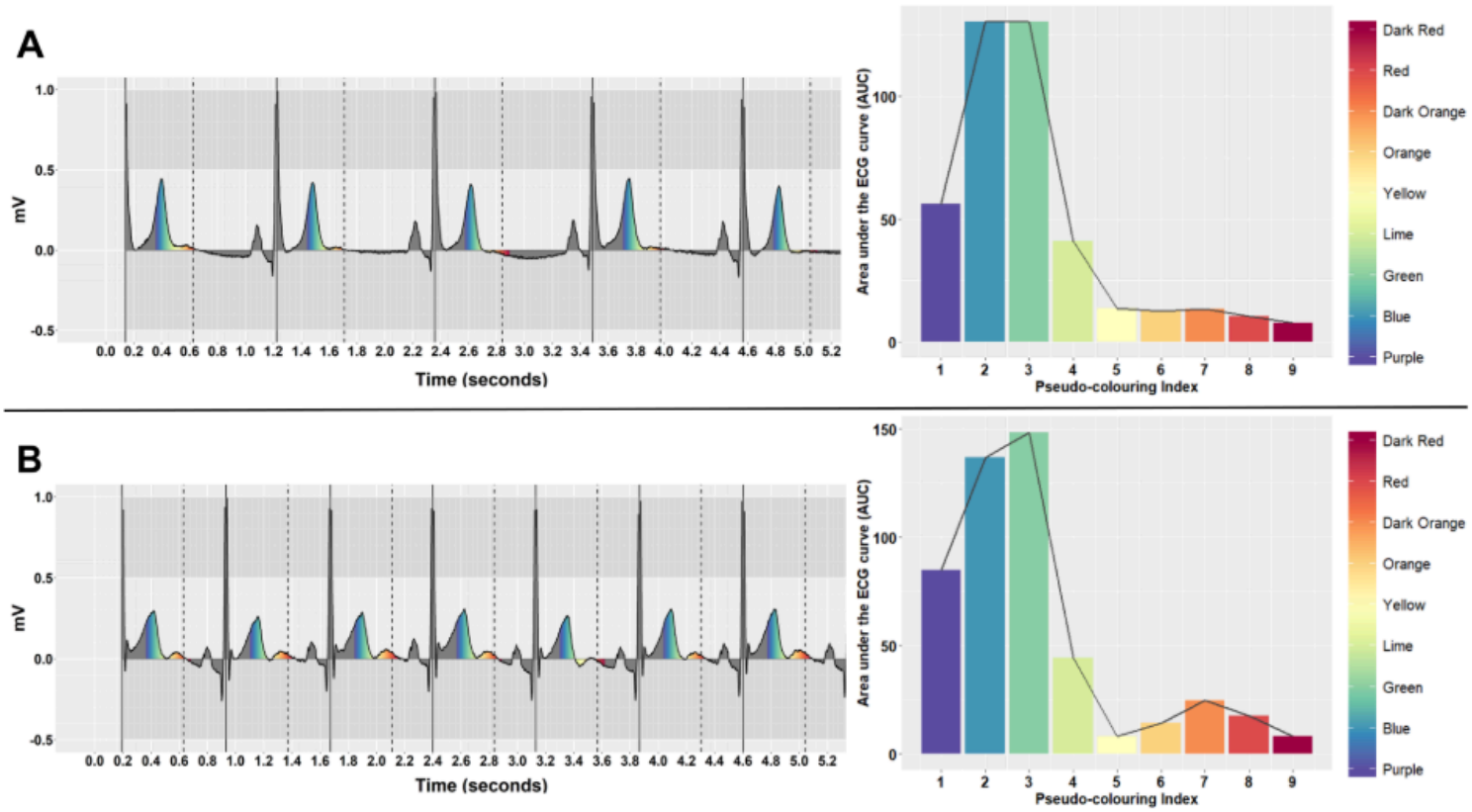


Figure 1

Pseudo-coloured ECGs with normal QT intervals (no risk of TdP) and their global representation graphs. In each panel, the T-wave area of each beat is partitioned into the nine colour segments (left), and the ∑AUC of each pseudo-colour, calculated across the whole ECG signal, is shown on the right (pseudo-colouring index 1 = purple to 9 = dark red). (A) An ECG with a low heart rate (HR = 55 bpm, QT = 361 ms) that displays a greater area of cool pseudo-colours; the small negative red area below the ECG baseline was included in the AUC calculation. (B) An ECG with a high heart rate (HR = 83 bpm, QT = 329 ms), with a cool pseudo-coloured T wave and a warm pseudo-coloured U wave. Reproduced from Fig. 6 of Alahmadi et al. [3] under a Creative Commons Attribution 4.0 (CC BY 4.0) licence

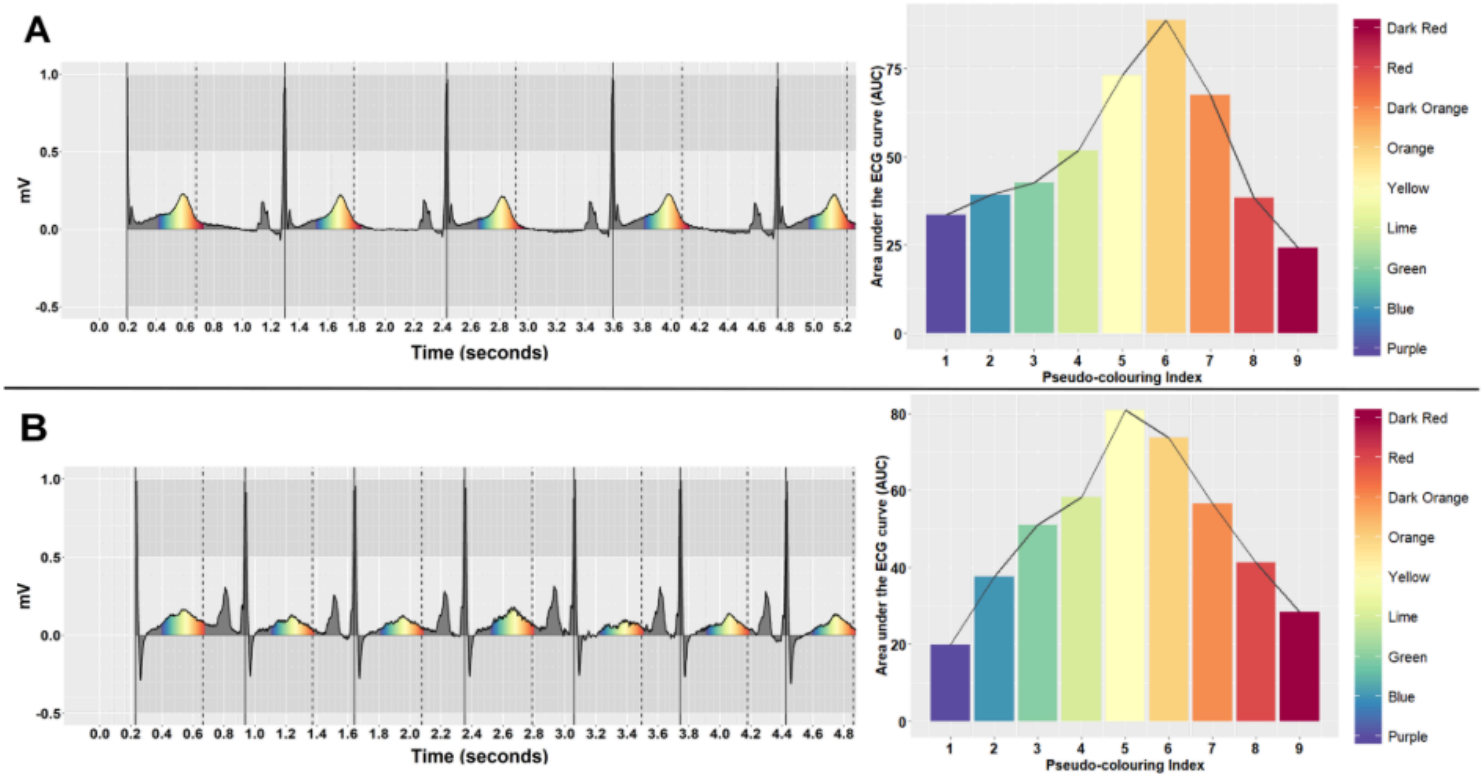


Figure 2

Pseudo-coloured ECGs with prolonged QT intervals (at risk of TdP) and their global representation graphs, showing the ∑AUC of each pseudo-colour calculated across the whole ECG signal (layout as in Fig. 1). (A) An ECG with a low heart rate (HR = 52 bpm, QT = 579 ms) that displays a greater area of warm pseudo-colours. (B) An ECG with a high heart rate (HR = 85 bpm, QT = 470 ms) and a notched T wave that contains a greater area of warm pseudo-colours. Reproduced from Fig. 7 of Alahmadi et al. [3] under a Creative Commons Attribution 4.0 (CC BY 4.0) licence

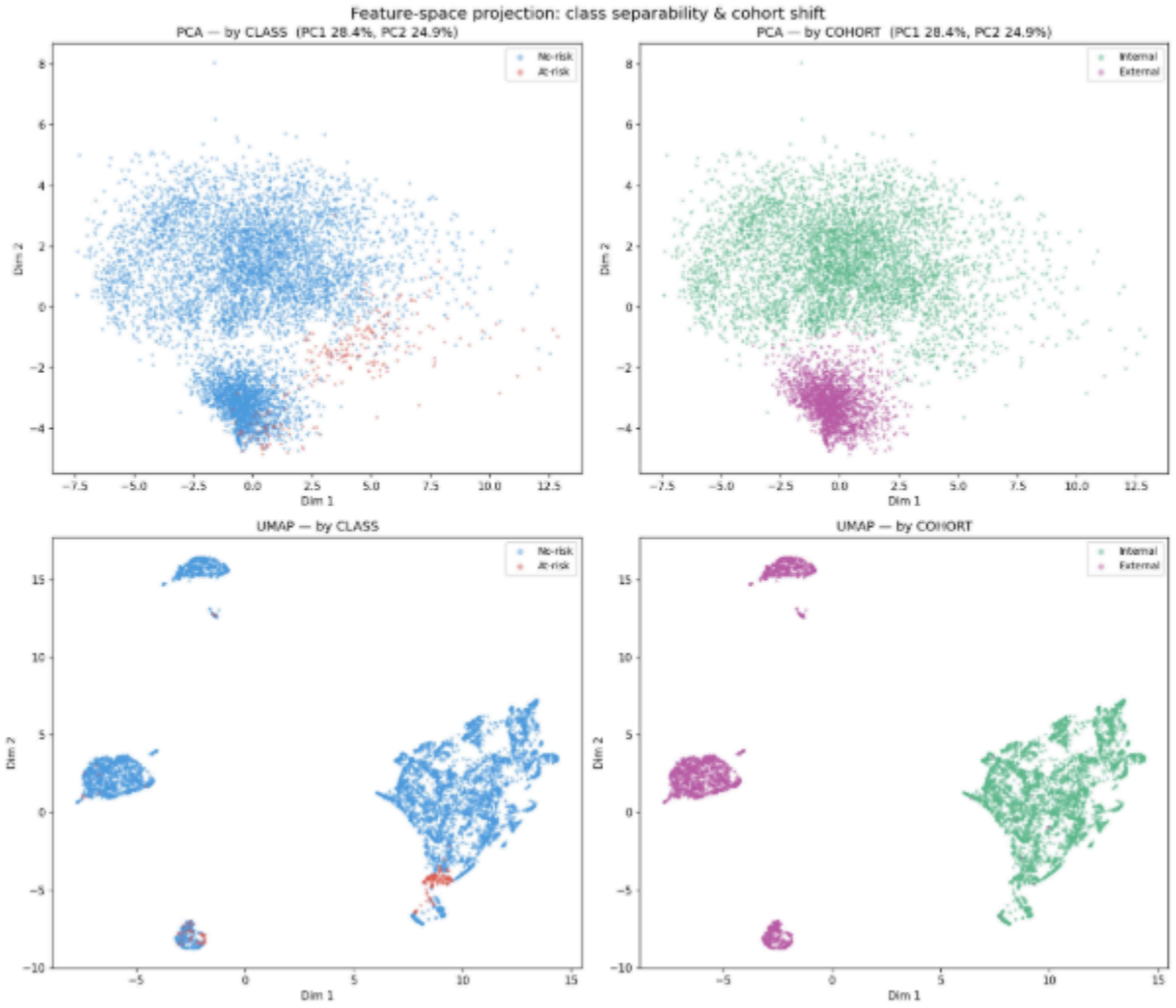


Figure 3

PCA and UMAP projections of the combined feature space, coloured by class and by cohort. The left panels show substantial overlaps between At-Risk and No-Risk samples, while the right panels show clearer separation between the internal and external cohorts, particularly in the UMAP projection

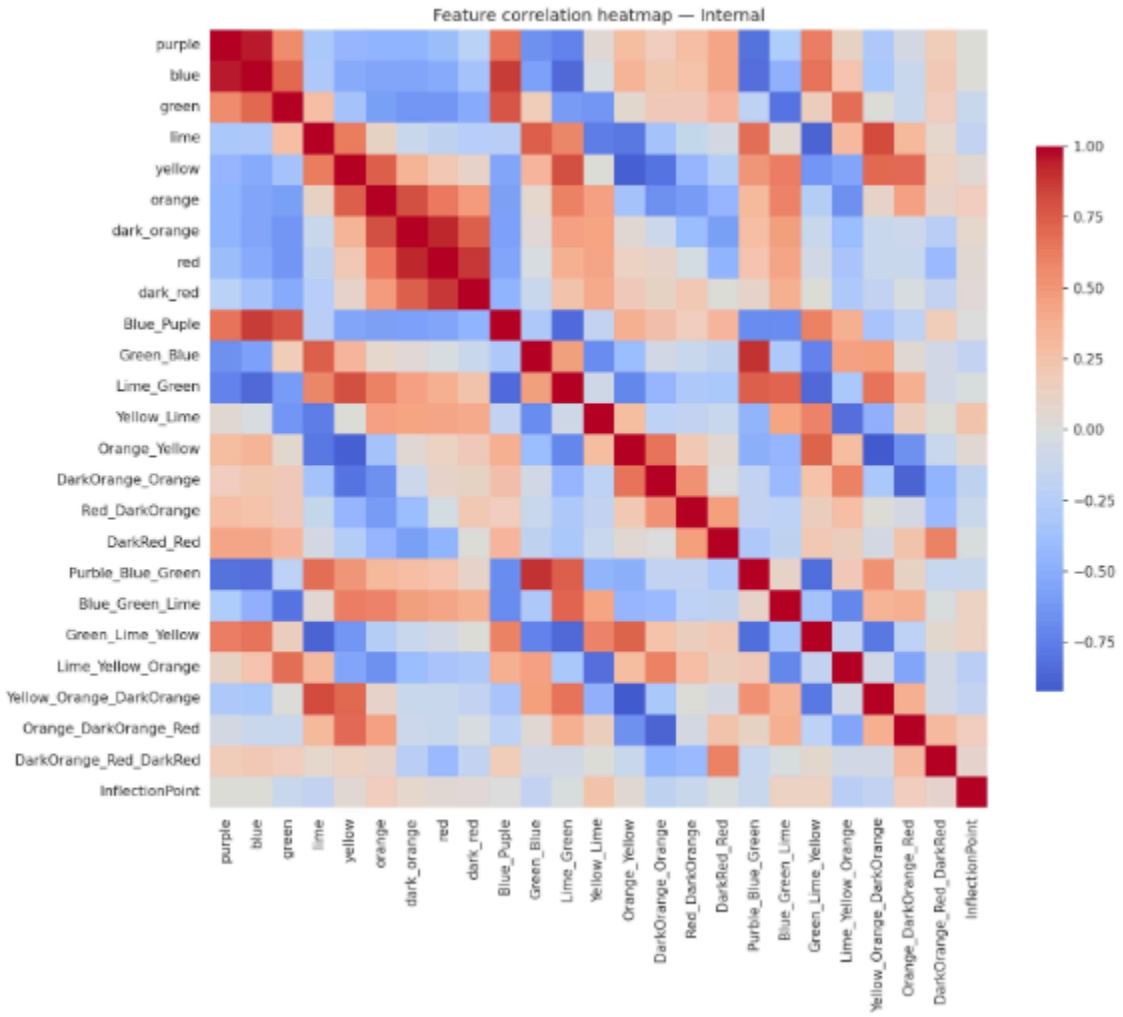


Figure 4

Pearson correlations between the 25 ECG-derived features in the internal cohort. Strong correlations are visible within the colour-based feature groups, particularly between neighbouring waveform segments, while InflectionPoint shows only weak correlations with most other features, indicating that it provides largely independent information

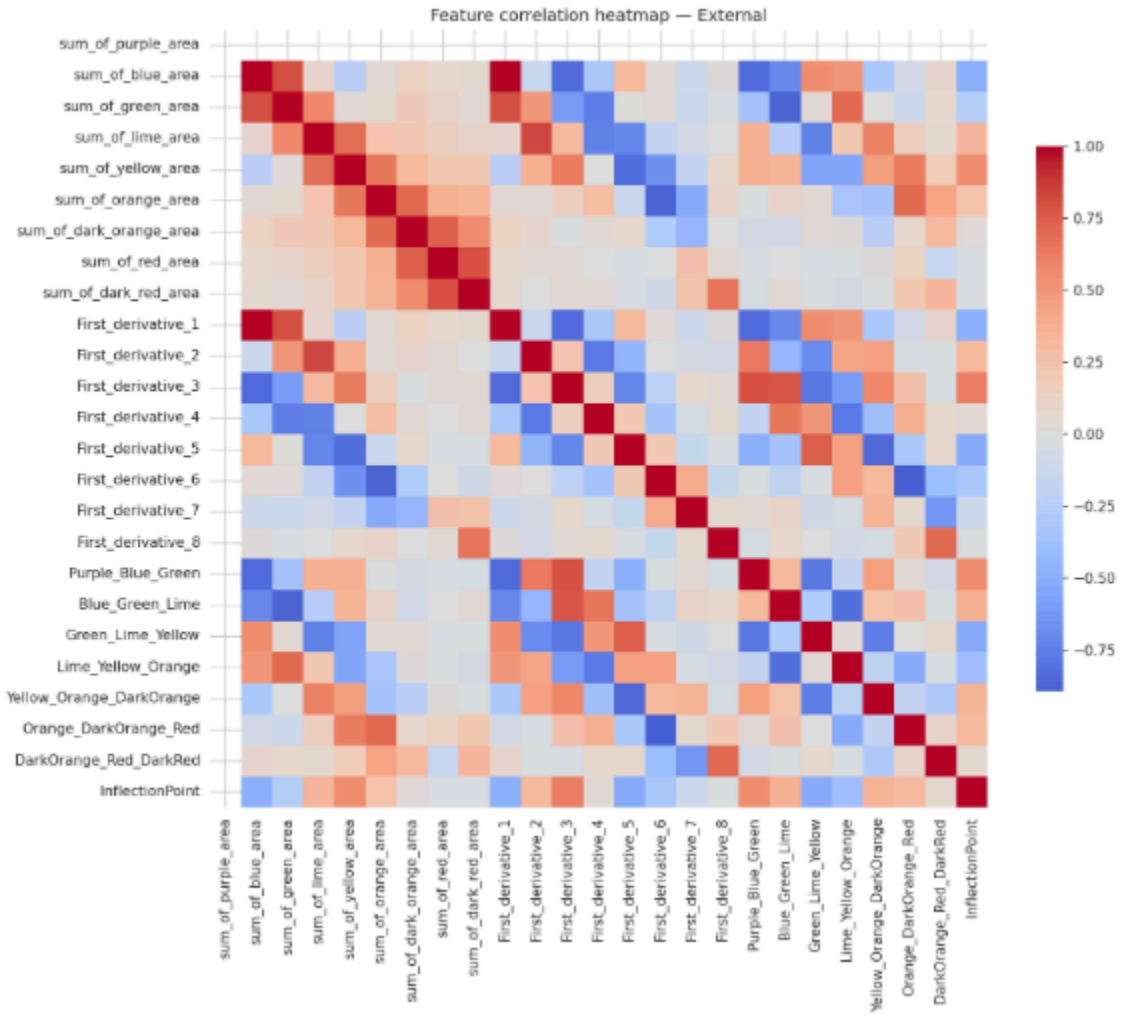


Figure 5

Pearson correlations between the 25 ECG-derived features in the external cohort. The structure is broadly similar to the internal cohort, with strong relationships among neighbouring colour-based features and weak correlations between InflectionPoint and most other variables. The purple feature has zero variance in this cohort, so its correlations are undefined and appear blank in the heatmap

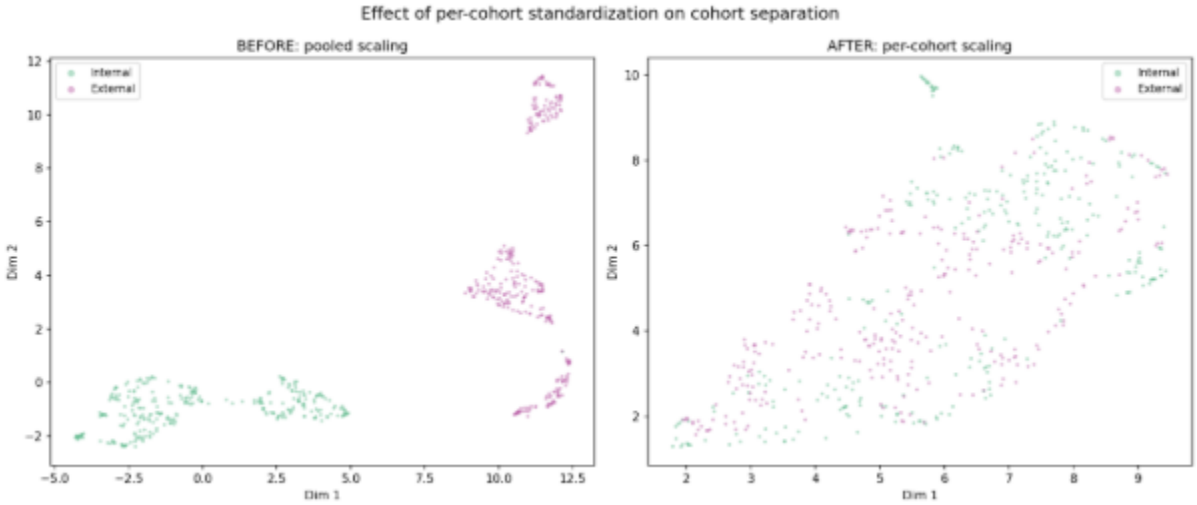


Figure 6

Two-dimensional embedding of the 25 ECG features for the internal (green) and external (pink) cohorts. Left: standardisation based on the internal data, as used in the external validation. Right: each cohort standardised with its own mean and variance, shown only as a diagnostic

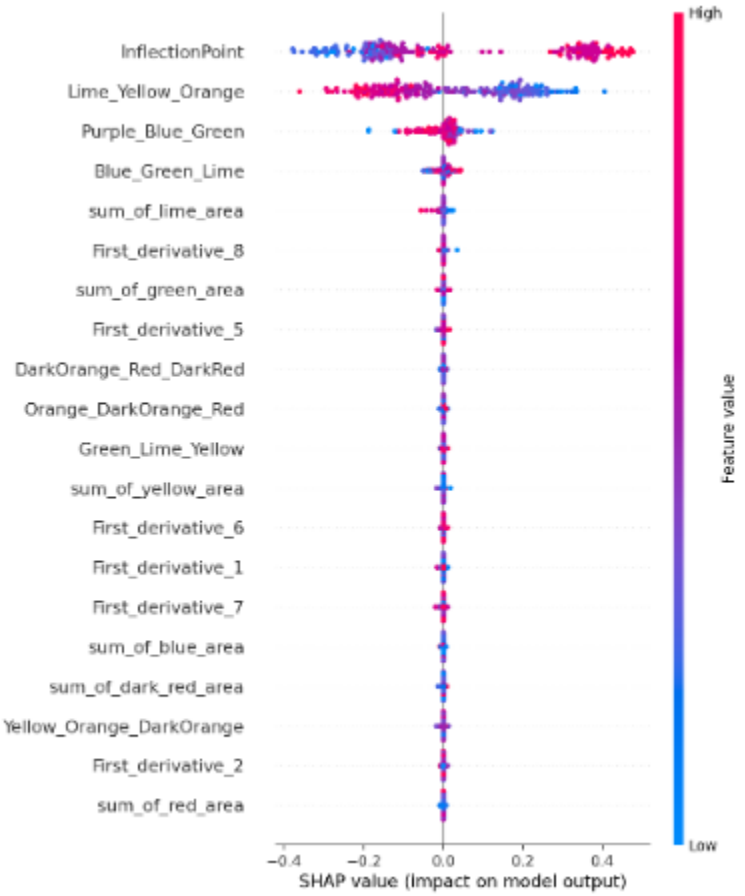


Figure 7

Kernel SHAP summary plot for the combined-cohort model (n = 600), ranking the 25 ECG-derived features by their mean absolute SHAP value (impact on the predicted probability of At-Risk status). Point colour indicates the corresponding feature value (blue = low, pink/red = high)

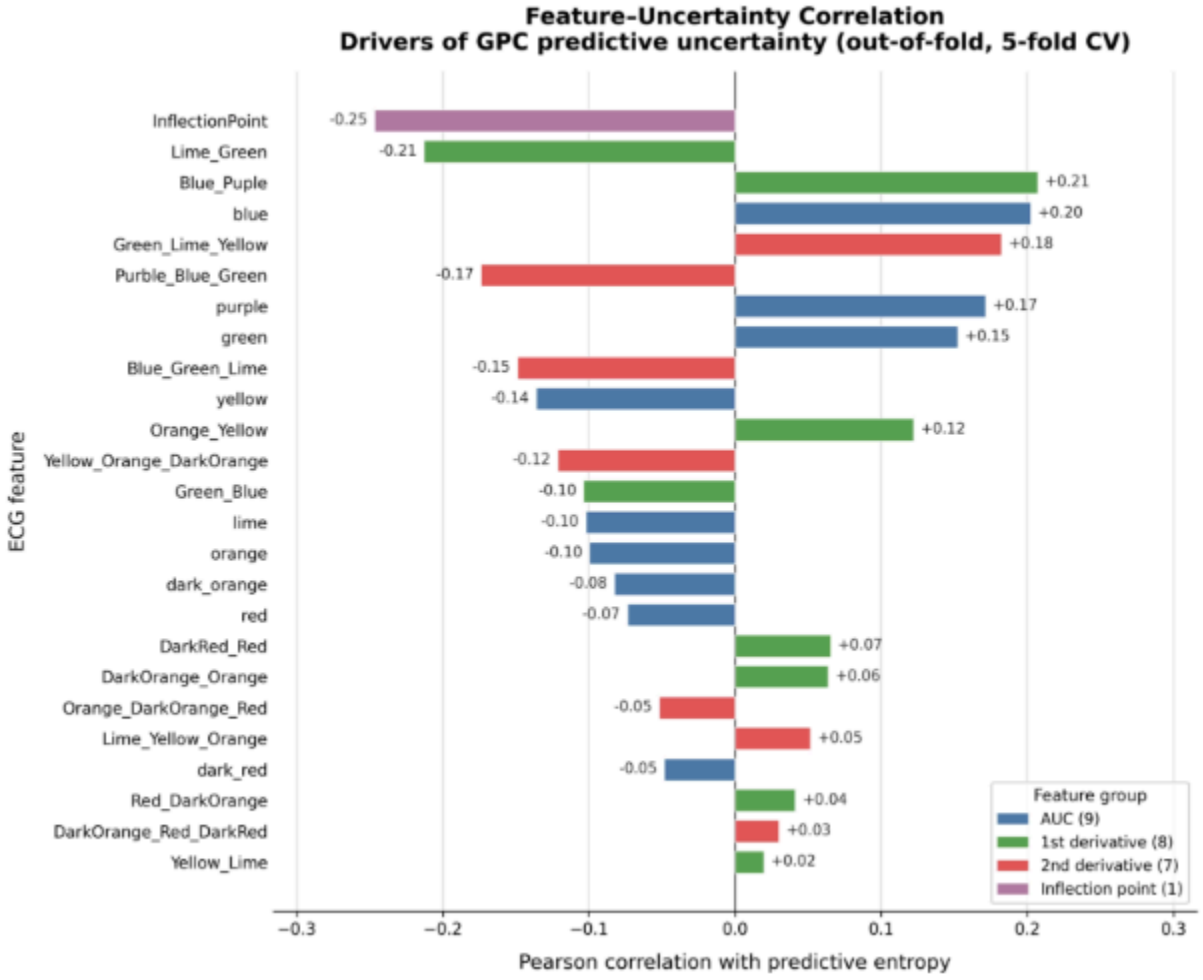


Figure 8

Pearson correlation between each of the 25 ECG-derived features and out-of-fold predictive entropy (uncertainty) of the GPC on the internal (ECGRDVQ) cohort (five-fold cross-validation), grouped by feature type (AUC, first derivative, second derivative, InflectionPoint)

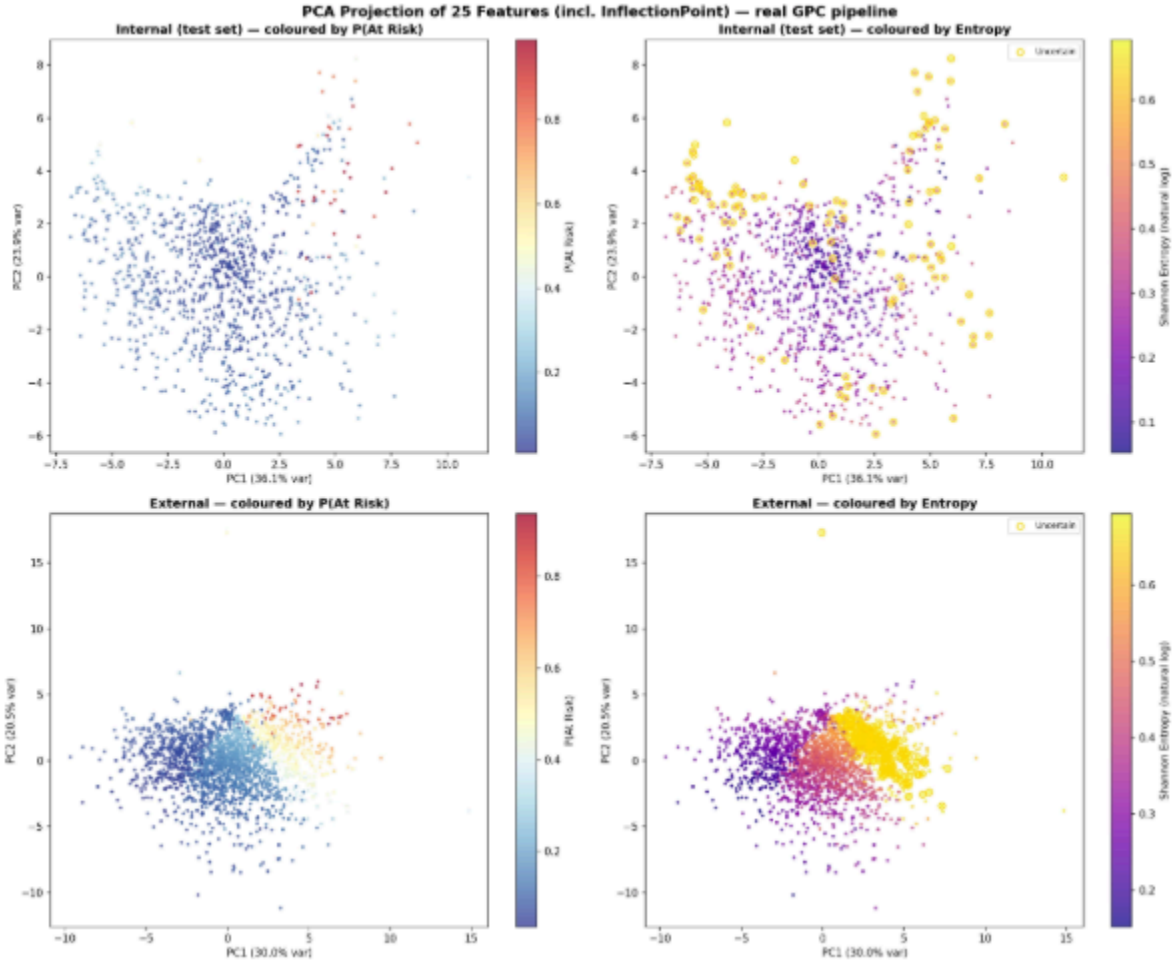


Figure 9

PCA projection of the 25 ECG-derived features for the internal and external cohorts. The left panels show the model-predicted probability of At-Risk status, while the right panels show predictive entropy

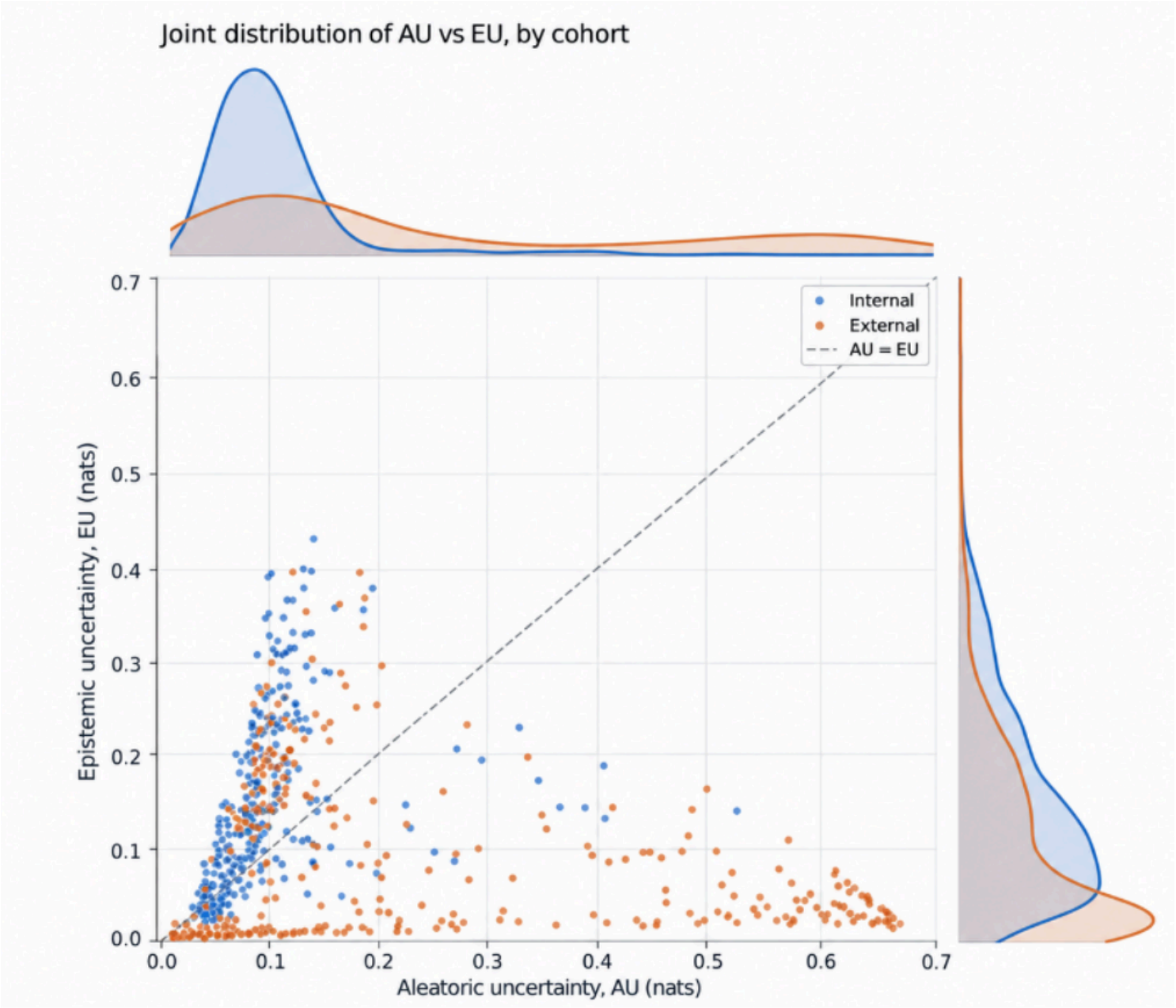


Figure 10

Distribution of aleatoric uncertainty (AU) and epistemic uncertainty (EU) in the combined dataset, separated by internal and external cohorts