Embracing Uncertainty Flexibility: Harnessing a Supervised Tree Kernel to Empower Ensemble Modelling for 2D Echocardiography-Based Prediction of Right Ventricular Volume

Tuan A. Bohoran [1,*,✉], Polydoros N. Kampaktsis[2,*], Laura McLaughlin[2],
Jay Leb[2], Serafeim Moustakidis[3], Gerry P. McCann[4], Archontis Giannakidis[1]

[1]School of Science and Technology, Nottingham Trent University, Nottingham, UK.;
Email: tuan.bohoran@ntu.ac.uk;
[2]Division of Cardiology, Columbia University Irving Medical Center, New York City, NY, USA.;
[3]AiDEAS, Tallinn, Estonia.;
[4]Department of Cardiovascular Sciences, University of Leicester and the NIHR Leicester Biomedical Research Centre, Glenfield Hospital, Leicester, UK.;
[*] Authors contributed equally

ABSTRACT

The right ventricular (RV) function deterioration strongly predicts clinical outcomes in numerous circumstances. To boost the clinical deployment of ensemble regression methods that quantify RV volumes using tabular data from the widely available two-dimensional echocardiography (2DE), we propose to complement the volume predictions with uncertainty scores. To this end, we employ an instance-based method which uses the learned tree structure to identify the nearest training samples to a target instance and then uses a number of distribution types to more flexibly model the output. The probabilistic and point-prediction performances of the proposed framework are evaluated on a relatively small-scale dataset, comprising 100 end-diastolic and end-systolic RV volumes. The reference values for point performance were obtained from MRI. The results demonstrate that our flexible approach yields improved probabilistic and point performances over other state-of-the art methods. The appropriateness of the proposed framework is showcased by providing exemplar cases. The estimated uncertainty embodies both aleatoric and epistemic types. This work aligns with trustworthy artificial intelligence since it can be used to enhance the decision-making process and reduce risks. The feature importance scores of our framework can be exploited to reduce the number of required 2DE views which could enhance the proposed pipeline's clinical application.

Keywords: uncertainty estimation, echocardiography, regression, machine learning, right ventricle, instance-based learning, ensemble models.

1. INTRODUCTION

Right ventricular systolic (RV) dysfunction is a powerful and independent mortality predictor [1] which may occur from a large variety of cardiovascular disorders that result in inability to pump enough blood for oxygenation. Machine learning methods have recently shown [2] great potential in quantifying RV volumes using tabular data (such as area measurements, age, gender and cardiac phase information) obtained from the widely available and highly portable two-dimensional echocardiography (2DE). However, for clinical deployment where patient safety is at stake, it is crucial to complement these RV volume predictions with uncertainty scores that reflect the degree of trust in these predictions.

The goal of this paper is to present an uncertainty quantification framework when predicting RV volumes through the use of ensemble models, in particular Gradient-Boosted Regression Trees (GBRTs), on 2DE-derived tabular data. GBRTs are regarded [3] the method of choice for tabular data. To get an estimate of the prediction uncertainty, we propose to make use of a k-nearest neighbour method [4] that relies on a supervised tree kernel [5, 6]. Unlike other state-of-the-art (SOTA) gradient-boosted algorithms [7–9] that provide probabilistic predictions, this method performs well on both probabilistic and point performances, and can also use a number of distribution types to more flexibly model the output. It can also be applied to any GBRT model, adding further flexibility.

The probabilistic and point-prediction performances of our framework are evaluated on a relatively small-scale dataset, comprising 100 end-diastolic (ED) and end-systolic (ES) RV volumes. The reference values for point performance were

obtained from cardiovascular MRI (CMR). Our pipeline is also compared to other SOTA methods. Lastly, we provide conditional output distributions and the respective confidence intervals for a couple of high and low accuracy (test set) predictions.

2. MATERIALS & METHODS

2.1 Dataset

The study population was a retrospective cohort of 50 adult patients for which 2DE and CMR were acquired. Data acquisition and annotation were as described in [2]. In brief, for each patient the RV endocardial-myocardial interface was manually traced in end-systole and end-diastole (making a total of 100 data points) for the following eight standardised echocardiographic views: parasternal long axis (PLAX), right ventricular inflow (RV Inflow), parasternal short axis at the level of the aortic valve (PSAX AV), basal (PSAX Base), mid (PSAX mid) and apical left ventricular segments (PSAX Distal), four-chamber (Four C) and subcostal (Sub C) views. The eight area measurements along with the patient age were the numerical input variables of our model, whereas the gender and cardiac phase information were the categorical ones. The short-axis cine CMR-derived ED and ES RV volumes were recorded in a semi-automated way and served as the reference values. The study was approved by the Columbia University Irving Medical Center Institutional Review Board and the Nottingham Trent University Ethics Committee.

2.2 Gradient-Boosted Regression Trees

Assume $D \coloneqq \{(x_i, y_i)\}_{i=1}^{n}$ is the training set where $x_i = (x_i^j)_{j=1}^{p} \in X \subseteq \mathbb{R}^p$ and $y_i \in Y \subset \mathbb{R}$. Gradient-boosting [10] is a powerful machine learning algorithm that constructs a model $f : X \to Y$ by relying on stage-wise additive modelling and minimising the expected value of some empirical loss function L. The model is obtained through the recursive relationship: $f_0(x) = \gamma, \ldots, f_t(x) = f_{t-1}(x) + \eta \cdot m_t(x)$. In this formulation, f_0 is the base learner, γ is an initial approximation, f_t is the model at iteration t, m_t represents the weak learner added during iteration t to boost the model, and η is the learning rate.

In the case of GBRTs, the most frequently employed L is the mean squared error (MSE), γ is chosen as the average outcome of the training instances ($\frac{1}{n}\sum_{i=1}^{n} y_i$), and regression trees represent the weak learners. The decision tree at iteration t is chosen to approximate the residual (or else the negative derivative of the loss function with respect to $\hat{y_i} = f_{t-1}(x_i)$), $m_t = \arg\min_m \frac{1}{n}\sum_{i=1}^{n}(-g_{i,t} - m(x_i))^2$, where $g_{i,t} = \frac{\partial L(y_i, \hat{y_i})}{\partial \hat{y_i}}$ is the functional gradient of the i-th training instance at iteration t. The decision tree at iteration t recursively creates M_t disjoint regions $\{r_j^t\}_{j=1}^{M_t}$ through partitioning the instance space. Each of these regions is termed a leaf. The parameter value θ_t^j for leaf j at tree t is commonly determined (given a fixed structure) through a one-step Newton method: $\theta_t^j = -\frac{\sum_{i \in I_t^j} g_t^i}{(\sum_{i \in I_t^j} h_t^i + \lambda)}$, where $I_t^j = \{(x_i, y_i) \mid x_i \in r_j^t\}_{i=1}^{n}$ is the instance set of leaf j for tree t, h_i^t is the second derivative of the i-th training instance with respect to $\hat{y_i}$, and λ acts as a regularization parameter. Hence, m_t can be denoted as: $m_t(x) = \sum_{j=1}^{M_t} \theta_t^j \mathbb{1}[x \in r_t^j]$ where $\mathbb{1}$ is the indicator function. Lastly, to generate a prediction for a target sample x_{te}, the final GBRT model sums up the values of the leaves traversed by x_{te} over all T iterations: $\hat{y}_{te} = \sum_{t=1}^{T} m_t(x_{te})$.

2.3 Instance-based Uncertainty Quantification

The goal is to estimate the conditional probability distribution $P(y|x)$ for some target variable y given some input vector x. To allow probabilistic predictions for any GBRT point predictor, we propose to make use of a method that adopts ideas from instance-based learning [11] and a supervised tree kernel, namely the Instance-Based Uncertainty quantification for GBRTs (IBUG) method [4]. To start with, the method capitalises on the fact that GBRTs yield accurate point predictions and uses this scalar output to model the conditional mean in a probabilistic forecast. Next, to further model the conditional output distribution, IBUG uses a supervised tree kernel [12] to more effectively identify the k training examples with the largest affinity to the target example. In particular, the affinity of the i-th training example x_i to a target example x_{te} is given by

$$A(x_i, x_{te}) = \sum_{i=1}^{T} \mathbb{1}[R_t(x_i) = R_t(x_{te})] \tag{1}$$

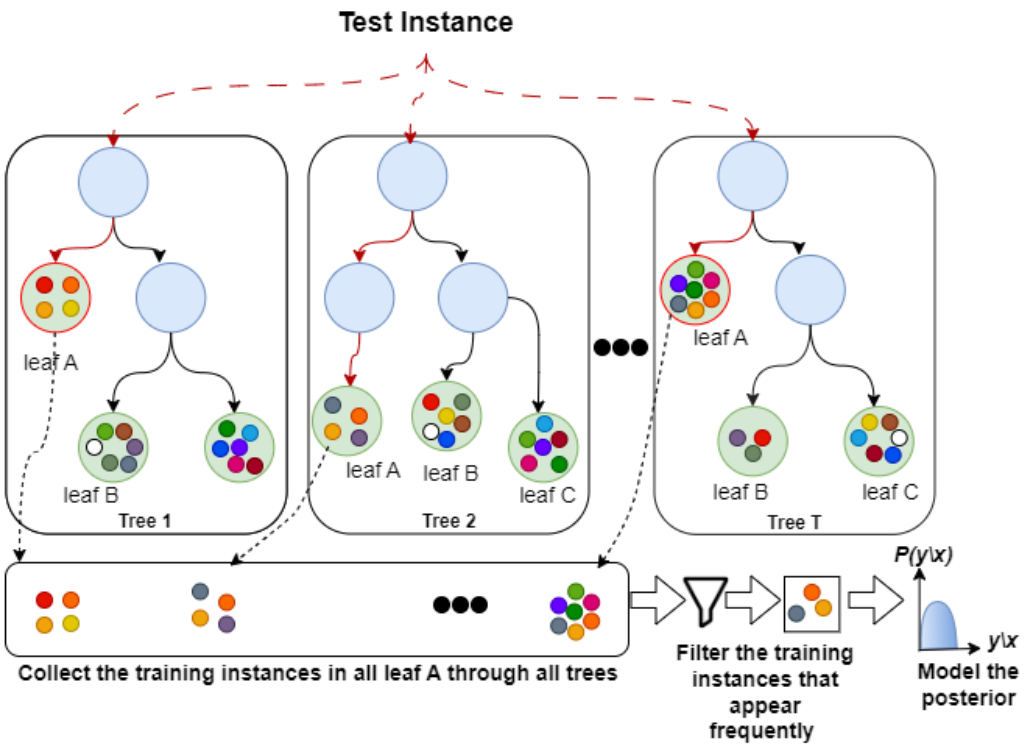


Figure 1: IBUG flow chart. For a target instance, IBUG collects the training instances at each leaf it traverses, keeps the k most frequent samples, and then uses those instances to model the output distribution.

where $R_t(x_i)$ is the leaf (of the tree t) to which x_i is assigned. Such a metric uses the structure of the learned trees in the ensemble. Lastly, the method employs the set of those k affinity scores, $A^{(k)}$, to produce a probabilistic prediction. The overall IBUG workflow is illustrated in Fig. 1.

Unlike other SOTA methods, IBUG offers numerous choices for modelling the conditional output distribution; the simplest way is, of course, through a normal distribution. In this case, the scalar output of the GBRT model is used to approximate the conditional mean ($\mu_{\hat{y}_{te}} = f(x_{te})$), and then the set $A^{(k)}$ is manipulated to compute the variance $\sigma^2_{\hat{y}_{te}}$. To further optimise the calculation of the prediction variance, the following calibration

$$\sigma^2_{\hat{y}_{te}} \leftarrow \gamma\sigma^2_{\hat{y}_{te}} + \delta \tag{2}$$

is commonly applied, where γ and δ are tuned on the validation set after the choice of k has been made. The method acts as a wrapper around any GBRT model allowing one to try various GBRT point predictors and then select the model with the best performance. To more flexibly model the output distribution using any parametric or non-parametric distribution, IBUG can use the set $A^{(k)}$ to directly fit (using maximum likelihood estimation) any continuous distribution D, including those with high-order moments:

$$\hat{D}_{te} = D\left(A^{(k)} \mid \mu_{\hat{y}_{te}}, \sigma^2_{\hat{y}_{te}}\right). \tag{3}$$

Choosing an appropriate value for k is critical for producing accurate probabilistic predictions. In this study, the tuning of k was performed in a held-out validation dataset $D_{\text{val}} \subset D$ using the negative log likelihood (NLL) probabilistic scoring metric. To accelerate the tuning process, through avoiding the repetition of the computationally expensive affinity calculations, the procedure described in Algorithm 3 of [4] was adopted, where parameter ρ was used to model instances of abnormally low variance.

2.4 Implementation

All implementations were in Cython. The experiments were conducted utilising an Intel Core i9 CPU 10900K Comet Lake, 10 Cores, 20 Threads @ 5.3GHz system equipped with 128GB of DDR4 RAM operating @ 2.6GHz. IBUG was

applied to XGBoost [13], LightGBM [14], and CatBoost [15]. We tuned k, using values: [3, 5, 7, 9, 11, 15, 31, 61, 91, 121, 151, 201, 301, 401, 501, 601, 701]. The parameters γ and δ were tuned using values ranging from 1×10^{-8} to 1×10^{3} with additional multipliers $[1.0, 2.5, 5.0]$. The number of trees, T, was tuned using values $[10, 25, 50, 100, 250, 500, 1000, 2000]$ (early stopping [7] was used for NGBoost). The learning rate was tuned using values $[0.01, 0.1]$. We also optimised: the maximum number of leaves h using values $[15, 31, 61, 91]$, the minimum number of leaves using values $[1, 20]$, and the maximum depth d using values $[2, 3, 5, 7, -1]$ (indicating no limit). The p parameter was adjusted based on the minimum variance obtained from the validation set predictions. We employed 5-fold cross-validation to generate 5 different 80/20 train/test folds. For each fold, the 80% training set was randomly divided into a 60/20 train/validation set for hyperparameter tuning. Upon tuning the hyperparameters, the model was retrained using the complete 80% training set. To test IBUG's flexibility in posterior modelling, we modelled each probabilistic prediction using the following distributions: normal, skewnormal, lognormal, Laplace, student t, logistic, Gumbel, Weibull, and KDE.

2.5 Evaluation Metrics

To evaluate probabilistic performance, the continuous ranking probability score (CRPS), NLL, check score, and interval score were used [16]. For all metrics, the lower the better. To gauge point performance, the root mean squired error (RMSE), the mean absolute error (MAE), the mean absolute percentage error (MAPE), the R^2 measure, and correlation were used. IBUG was compared to three recent gradient boosting algorithms that provide probabilistic predictions, namely NGBoost [7], PGBM [8], and CatBoost with uncertainty (CBU) [9].

2.6 Exemplars

To facilitate a comprehensive understanding of our results, we provide conditional output distributions and the respective confidence intervals (CIs) for two high and two low accuracy (test set) predictions.

2.7 Explainability

A benefit of using GBRTs is that it is straight-forward to retrieve feature importance scores that indicate how valuable each attribute was in the construction of the model. In this study, feature importance scores were calculated for all models using the 'Gain' metric which quantifies the relative contributions.

3. RESULTS

The final set of hyperparameters for each method and the corresponding tuning and training times are listed in Table 1. Table 2 compares the probabilistic performance of the IBUG model against the three SOTA probabilistic prediction methods. IBUG model with CatBoost as the base learner provided the lowest average scores in all CRPS, NLL, Check Score and Interval Score indices. In Table 3, the point performance of all methods is provided. In overall, IBUG method with CatBoost base learner displayed the best performance once again. Table 4 shows the importance of variance calibration in the probabilistic performance of IBUG. Table 5 demonstrates that the logistic (parametric) distribution better fits the underlying data than assuming normality. In Fig. 2 and Fig. 3, we illustrate the conditional output distributions for four representative test cases (two that were predicted with high accuracy and two that were predicted with low accuracy), when normal and logistic probabilistic density functions were used for modelling, respectively. Table 6 lists the 95% and 99% confidence intervals for the above cases. These results showcase the appropriateness of the proposed framework for providing uncertainty scores for RV volume predictions. Lastly, in Fig. 4, we illustrate the "Gain" feature importance score for all eleven features in the best IBUG model. The parasternal long axis (PLAX), four chambers (Four C) and parasternal short axis at base level (PSAX Base) standard views were the top three contributors to the model predictions.

Table 1: The final set of hyperparameters used for each method. Also shown are the tuning and training times. IBUG was applied to CatBoost, XGBoost, and LightGBM.

Parameter	CatBoost	XGBoost	LightGBM	NGBoost	PGBM	CBU
K	5	15	3	-	-	-
δ	1	0.5	0.5	5	10	1
Operation	add	mult	mult	add	add	add
min scale	6.164	13.826	2.055	-	-	-
n estimators (trees)	100	25	25	244	250	250
maximum depth	5	2	-1	-	-	-
learning rate	0.1	0.1	0.1	0.01	0.01	0.1
minimum data in_leafv	1	-	-	-	20	1
minimum child weight	-	20	20	-	-	-
number of leaves	-	-	15	-	15	15
max bin	255	255	255	255	255	255
tune+train time (s)	67.369	19.932	14.260	5.370	872.663	81.929

Table 2: Probabilistic performance comparison on the test set (five folds). IBUG results are for the case when CatBoost was the base learner. IBUG results have been averaged over all nine posterior output distributions.

Method	NLL	CRPS	Check Score	Interval Score
IBUG	**4.747**	**15.398**	**7.775**	**73.380**
NGBoost	7.571	22.174	11.177	141.618
PGBM	6.136	20.796	10.492	122.401
CBU	5.780	19.524	9.853	110.140

Table 3: Point performance comparison on the test set (five folds). IBUG results are for the case when CatBoost was the base learner.

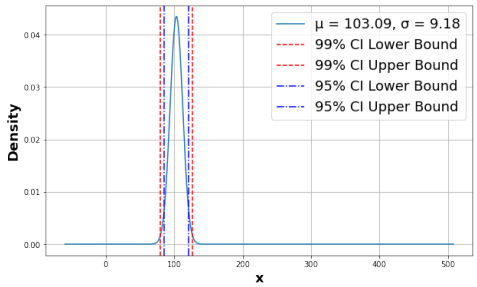
Method	MAE	RMSE	MAPE	R^2	Correlation
IBUG	**22.75**	**26.292**	**20.22**	**0.666**	**0.824**
NGBoost	28.114	32.269	24.406	0.496	0.736
PGBM	27.479	31.27	25.147	0.527	0.768
CBU	26.378	30.127	22.974	0.561	0.772

Table 4: Probabilistic performance comparison of IBUG method with and without variance calibration. IBUG results are for the case when CatBoost was the base learner.

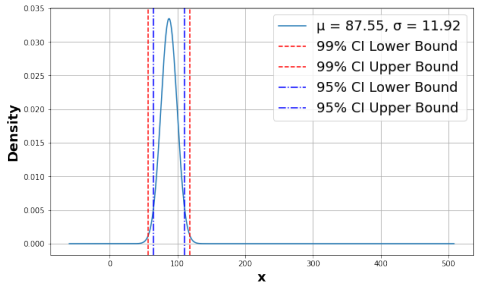
Operation	NLL	CRPS	Check Score	Interval Score
With Calibration	**4.747**	**15.398**	**7.775**	**73.38**
Without Calibration	4.781	15.457	7.805	74.044

Table 5: Probabilistic performance comparison when assuming normal and logistic distributions for modelling the underlying data.

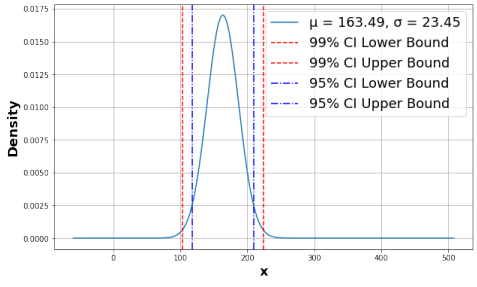
Distribution	NLL
Normal	5.10466
Logistic	**5.00837**



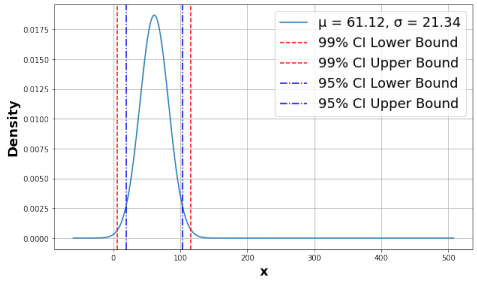
(a)



(b)

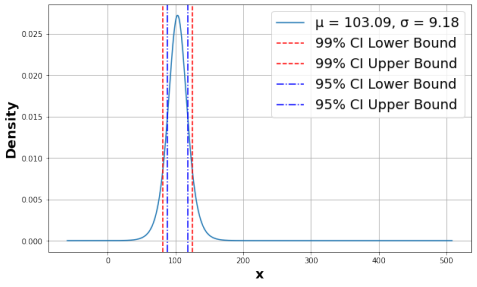


(c)

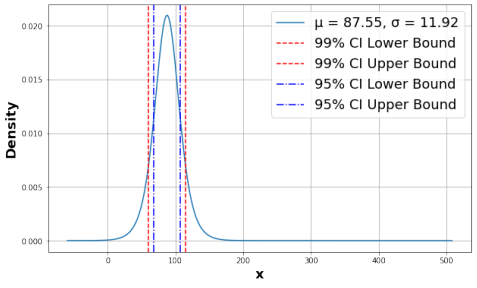


(d)

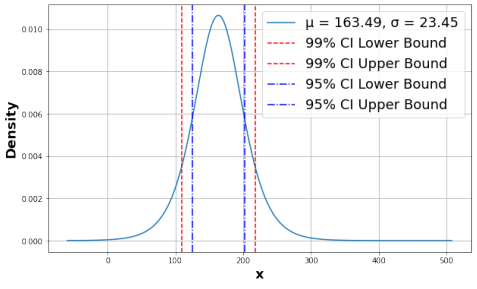
Figure 2: The conditional output normal distributions for test instances that were predicted with high [(a) and (b)] and low [(c) and (d)] accuracy.



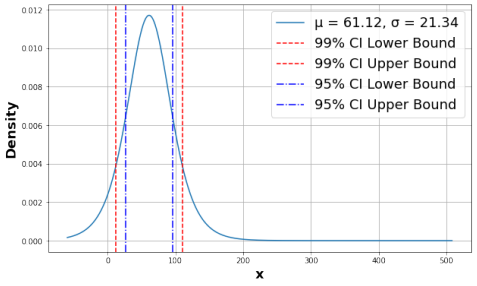
(a)



(b)



(c)



(d)

Figure 3: The conditional output logistic distributions for test instances that were predicted with high [(a) and (b)] and low [(c) and (d)] accuracy.

Table 6: The 95% and 99% Confidence Intervals for both Normal and Logistic distributions for four representative test set cases, two that were predicted with high accuracy (low APE) and two that were predicted with low accuracy (high APE).

Prediction Accuracy	APE (%)	Point Prediction	Normal Distribution		Logistic Distribution	
			95% CI	99% CI	95% CI	99% CI
High	3.090	103.090	35.979	47.286	30.196	42.697
	0.508	87.553	46.739	61.428	39.227	55.466
Low	10.169	163.492	91.932	120.825	77.157	109.099
	10.119	61.119	83.650	109.939	70.206	99.270

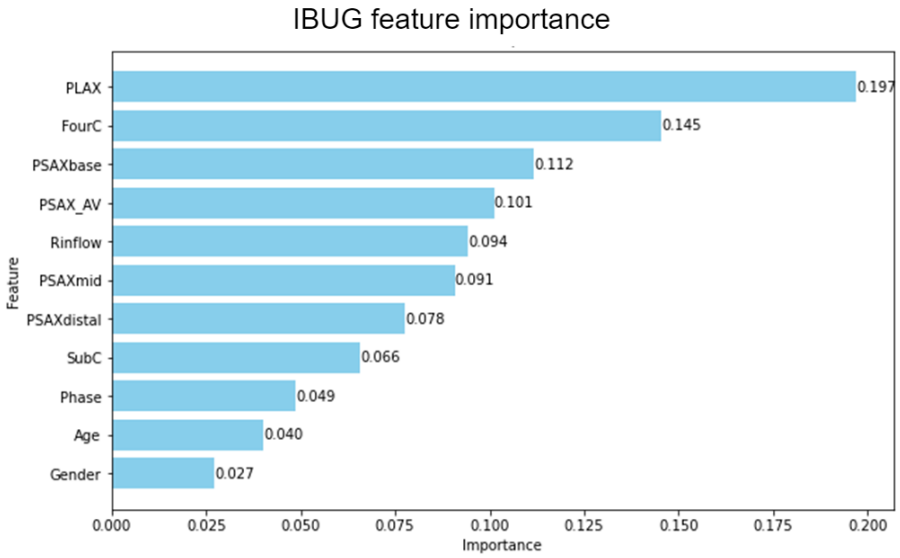


Figure 4: Feature importance plot for the IBUG model with CatBoost as the base learner.

4. DISCUSSION AND CONCLUSIONS

In this paper, we presented an uncertainty quantification framework when predicting RV volumes through the use of GBRTs on 2DE tabular data (such as area measurements, age, gender and cardiac phase information). To get an estimate of the prediction uncertainty, we employed the IBUG method which uses the learned tree structure to identify the k nearest training samples to a target instance. The results on a small-scale dataset demonstrate that this simple wrapper yields improved probabilistic and point performances over other SOTA methods. The appropriateness of the proposed framework for providing uncertainty scores for RV volume predictions was showcased by providing conditional output distributions and confidence intervals for four exemplar cases. Additional research is required, involving a larger sample size of patients and encompassing a broader range of RV volumes, to substantiate these findings. The estimated uncertainty embodies both aleatoric and epistemic types of uncertainty since IBUG is an instance-based approach and also predictions on the training set were used to tune k, γ, and δ. Overfitting was observed in IBUG's point performance which is a typical finding when the size of the dataset is small. This work aligns with trustworthy artificial intelligence [17] since it can be used to enhance the decision-making process and reduce risks. It could help overcome mistrust which is a major barrier to the deployment of machine learning systems in the clinical setting. The calculated feature importance scores can be used for reducing the number of required 2DE views, which in turn could enhance the proposed pipeline's clinical application.

ACKNOWLEDGMENTS

Funding

Tuan Aqeel Bohoran is funded by the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 801604.

REFERENCES

[1] F. Haddad, S. Hunt, D. Rosenthal, and D. Murphy, "Right ventricular function in cardiovascular disease, part i: Anatomy, physiology, aging, and functional assessment of the right ventricle.," Circulation **117**, 1436–48 (2008). 2008,3,18.

[2] T. Bohoran, P. Kampaktsis, L. McLaughlin, J. Leb, S. Moustakidis, G. McCann, and A. Giannakidis, "Right ventricular volume prediction by feature tokenizer transformer-based regression of 2d echocardiography small-scale tabular data," in Functional Imaging And Modeling Of The Heart, 292–300 (2023).

[3] L. Grinsztajn, E. Oyallon, and G. Varoquaux, "Why do tree-based models still outperform deep learning on typical tabular data?," in Advances In Neural Information Processing Systems, **35**, 507–520 (2022).

[4] J. Brophy and D. Lowd, "Instance-based uncertainty estimation for gradient-boosted regression trees," in Advances In Neural Information Processing Systems, **35**, 11145–11159 (2022).

[5] Y. Lin and Y. Jeon, "Random forests and adaptive nearest neighbors," Journal of the American Statistical Association **101**(474), 578–590 (2006).

[6] T. Daghistani and R. Alshammari, "Comparison of statistical logistic regression and randomforest machine learning techniques in predicting diabetes," Journal of Advances in Information Technology **11**, 78–83 (may 2020).

[7] T. Duan, A. Anand, D. Ding, K. Thai, S. Basu, A. Ng, and A. Schuler, "Ngboost: Natural gradient boosting for probabilistic prediction," in Proceedings Of The 37th International Conference On Machine Learning, **119**, 2690–2700 (2020). 2020,7,13.

[8] O. Sprangers, S. Schelter, and M. Rijke, "Probabilistic gradient boosting machines for large-scale probabilistic regression," in Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, (2021).

[9] A. Malinin, L. Prokhorenkova, and A. Ustimenko, "Uncertainty in gradient boosting via ensembles," in International Conference On Learning Representations, (2021).

[10] J. Friedman, "Greedy function approximation: A gradient boosting machine," The Annals Of Statistics **29**, 1189–1232 (2001).

[11] D. Aha, D. Kibler, and M. Albert, "Instance-based learning algorithms," Machine Learning **6**, 37–66 (1991).

[12] A. Davies and Z. Ghahramani, "The random forest kernel and other kernels for big data from random partitions," arXiv preprint arXiv:1402.4293 (2014).

[13] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," CoRR **abs/1603.02754** (2016).

[14] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," in Advances In Neural Information Processing Systems, **30** (2017).

[15] L. Prokhorenkova, G. Gusev, A. Vorobev, A. Dorogush, and A. Gulin, "Catboost: unbiased boosting with categorical features," in Advances In Neural Information Processing Systems, **31** (2018).

[16] T. Gneiting and A. Raftery, "Strictly proper scoring rules, prediction, and estimation," Journal Of The American Statistical Association **102**, 359–378 (2007).

[17] B. Li, P. Qi, B. Liu, S. Di, J. Liu, J. Pei, J. Yi, and B. Zhou, "Trustworthy ai: From principles to practices," ACM Comput. Surv. **55** (2023). 2023,1.