**Machine Learning Model for predicting DIBH non-eligibility in Left-Sided Breast Cancer Radiotherapy: Development, Validation and Clinical Impact Analysis**

**ML model predicts DIBH ineligibility in breast RT**

**Authors:**

Dr Kundan Singh Chufal, MD 1

Dr Irfan Ahmad, DNB 1

Dr Alexis Andrew Miller, FRANZCR 2

Dr Ram Bajpai, PhD 3

Ms Avani Dwivedi 4

Mr Alok Dwivedi, MBA 5

Dr Preetha Umesh, DNB 1

Dr Kratika Bhatia, DNB 1

Dr Munish Gairola, MD 1

**Author Affiliations:**

1 Department of Radiation Oncology, Rajiv Gandhi Cancer Institute & Research Centre, New Delhi, India

2 Department of Radiation Oncology, Illawarra Cancer Care Centre, Wollongong, New South Wales, Australia

3 School of Medicine, Keele University, Staffordshire, United Kingdom

4 Department of Computer Science, Royal Holloway University of London, United Kingdom

5 Discover Financial Services, Reading, United Kingdom

**Corresponding Author:**

Dr Irfan Ahmad,

Department of Radiation Oncology, Rajiv Gandhi Cancer Institute & Research Centre,

Sector 5, Rohini, New Delhi, India

Email, Phone Number: irfan.a@icloud.com, (0091)-11-47022012

**Authors Responsible for Statistical Analysis:**

1. Dr Kundan Singh Chufal, Email: kundan25@gmail.com

2. Dr Irfan Ahmad, Email: irfan.a@icloud.com

3. Dr Ram Bajpai, Email: rambajpai@hotmail.com

**Human Ethics and Consent to Participate declarations:** This study’s protocol was submitted to and reviewed by the Institutional Review Board and Ethics Committee (IRB/IEC) of Rajiv Gandhi Cancer Institute & Research Centre, New Delhi, India (RGCIRC). After protocol review the IRB/IEC of RGCIRC approved the study and since all patients provided consent for DIBH assessment and treatment, the need for a separate consent for this study was waived (RES/SCM/62/2024/ 01). The IRB/IEC of RGCIRC is registered with Department of Health Research (EC/NEW/INST/2020/74).

**Conflicts of Interest:** None

**Availability of data and materials:** This study was performed at Rajiv Gandhi Cancer Institute & Research Centre and is stored in the institutional data repository. The authors do not own these data and hence are not permitted to share them in the original form (only in aggregate form). Reasonable requests for access to data will be considered on an individual basis, by contacting the corresponding author. The decision to share data will be taken by the members of the Data Monitoring & Ethics Committee.

**Funding:** The study was funded by Rajiv Gandhi Cancer Institute & Research Centre

**Authors’ Contributions:** All authors contributed equally in the design, conception and writing of this manuscript.

**Acknowledgements:** We are grateful to Dr Sudhir Rawal, Mr Rajesh Thacker & Mr DS Negi for their encouragement and support.

**ABSTRACT**

**Objective:** Multi-day assessments accurately identify patients with left-sided breast cancer who are ineligible for irradiation in Deep Inspiration Breath Hold (DIBH) and minimise on-couch treatment time in those who are eligible. The challenge of implementing multi-day assessments in resource-constrained settings motivated the development of a machine learning (ML) model using data only from the 1st day of assessment to predict DIBH ineligibility.

**Methods:** This prospective cohort study used data from 202 patients collected between January and December 2023 for model development. Patient-related and DIBH assessment-related variables (upper, lower, and average breath-hold amplitude; average breath-hold duration; breath-hold consistency) were included. Nine ML algorithms (and three modelling strategies) were evaluated, and a decision curve analysis was used to select the best model.

The best model was temporally validated on a prospective dataset of 47 patients (January to March 2024). Further, a clinical impact study on another prospective cohort of 64 patients (April to August 2024) was performed, to assess its practical utility by comparing its predictions with the clinical team’s decision to treat a patient in DIBH or not.

**Results:** The uncalibrated gradient-boosting ensemble model demonstrated the highest performance [AUC (95% CI) = 0.803 (0.686-0.941); Recall = 0.526] and net benefit in decision curve analysis. Key predictors included average breath-hold duration and lower breath-hold amplitude levels. The clinical impact study suggests that the model reduces the need for additional DIBH assessments by up to 20% without misclassifying eligible patients.

**Conclusion:** The developed ML model accurately predicts DIBH ineligibility using only first-day DIBH assessment data and could be a decision aid for patient selection in resource-constrained or busy departments. External validation is necessary to confirm its generalizability.

**Keywords:** Breast Neoplasms; Radiotherapy; Deep Inspiration Breath Hold; Machine Learning; Eligibility Determination

**Introduction**

Deep Inspiration Breath Hold (DIBH) radiotherapy is the standard of care for patients with left-sided breast cancer as it remarkably reduces cardiac exposure and consequently long-term cardiac morbidity.[1–4] Its acceptance by patients and radiation therapists, as well as its application in departments with a heavy workload has also been established.[5–7] While the principle of maintaining a moderately deep breath-hold is obvious, no formal definition of DIBH eligibility exists and since the clinical practice of DIBH is not standardised, limited guidance is available for new departments.[1,8,9]

Once a department begins planning a DIBH program, considerations include treating all patients in DIBH or pre-selecting them based on predicted cardiac dosimetric benefit.[2,10] Next, the duration of assessment/coaching to determine DIBH eligibility needs to be determined. Practices vary, and it is common for patients to be assessed on the day of CT simulation alone.[11] Alternatively, devoting more time to assessing/coaching patients improves breath-hold stability and reproducibility, thereby reducing on-couch time.[12,13] The decision to treat all with DIBH or pre-select them based on a finite resource (on-couch time or predicted dosimetric benefit) is strongly influenced by patient volume.

A ‘*default* DIBH’ approach benefits the maximum number of patients because rarely (if ever) does the cardiac exposure exceed that of a competing free-breathing RT plan.[14,15] Based on the no-threshold phenomenon of cardiac morbidity, proponents of this approach argue that any reduction in heart Dmean (MHD) is valuable.[16,17] Obviously, this strains available resources (both machine and personnel), especially in high-volume departments, for whom pre-selecting patients based on a predicted dosimetric benefit is a pragmatic approach, but requires an answer to the question, ‘What reduction in MHD is *not* potentially beneficial?’.[18–20] High-volume departments also need to consider on-couch time, which is correlated with a patient’s intrinsic ability to maintain a reproducible breath-hold and without a formal definition, whether a patient will be able to do so, becomes a subjective decision taken at the time of simulation.[1]

Variables influencing this decision include breath-hold reproducibility (inter- or intra-fraction variation), reproducible breath-hold duration (as shorter breath-holds can prolong on-couch time), and patient-specific factors such as age, BMI, and smoking history. The type of DIBH equipment (surface- or marker-based tracking, spirometry-based systems, or an equipment-free approach) and experience of the treating team (or lack thereof) are also contributory. Preliminary investigations suggest that objective data extracted from DIBH waveforms generated during three days of assessment are associated with determining eligibility.[21]

Lower-middle income countries (LMIC) face unique challenges (limited availability of DIBH-capable LINACs and trained personnel; increasing incidence of breast cancer and cardiovascular disease) which precludes treating all patients in DIBH. Accurately identifying patients who can receive treatment for 15-20 sessions with reproducible breath-hold will benefit the maximum number of patients in LMICs, without unduly prolonged on-couch times.[22,23]

To address this specific regional healthcare challenge, the primary objectives of this study were: (a) to develop and validate a machine learning model designed to identify patients who are unlikely to complete treatment in DIBH successfully (“ineligible”) based on clinical variables and data extracted from the first day of DIBH assessment; (b) to assess the clinical impact of this model before clinical deployment.

**Materials and Methods**

**A) Study design, data description, primary outcome definition and data processing**

This prospective, single institution, IRB-approved [protocol ID: RES/SCM/62/2024/01, IRB Approval ID: RGCIRC/IRB-BHR/30/2024] observational study was conducted between January 2023 and August 2024. It utilized a retrospective dataset for model development (January 2023 to December 2023) and a prospectively collected dataset for model evaluation (January 2024 to March 2024). Details of inclusion, exclusion criteria, and dataset description have been previously reported (please see study protocol in supplemental materials).[21] The clinical impact of the developed model was tested on another prospectively collected dataset (April 2024 to August 2024). Patients who declined or did not complete DIBH assessment were excluded from the study.

Clinical and DIBH parameters from the first day of assessment were retrieved from our workflow management system (OncFlowTM, Dashamlav AI Labs, India). Predictors were selected from a literature review and expert consensus. Patient-related factors were: age (years), body mass index (BMI), surgery type (breast conservation or mastectomy), chemotherapy type (no chemotherapy, adjuvant, or neoadjuvant), and comorbidities (none, cardiac comorbidities, or other comorbidities). DIBH assessment–related factors were: the upper, lower, and average levels of breath-hold amplitude (mm), average hold duration (seconds), and whether the patient demonstrated a consistent breath hold (yes or no). Please see [21] and supplemental materials for a detailed description of the predictors.

Continuous variables were independently standardized using *StandardScaler* (part of Python's SciKit-learn library), using the formula, z = (x - u) / s; where 'x' is the variable to be standardized, 'u' and 's' are the mean and standard deviation of the variable.[24] Formal guidance on sample size calculation when using ML models is absent.

The primary binary outcome was the DIBH eligibility decision taken by the treating radiation oncologist, therapist, and medical physicist team (for a detailed description of the procedure, please see[21]). The outcome was coded to identify patients ineligible for the DIBH technique (true positive). Therefore, a higher predicted probability in the model meant that a patient was more likely to be ineligible for DIBH. Conversely, a lower probability meant that a patient was more likely to be eligible for DIBH (true negative).

The ground truth was the successful completion of treatment using DIBH. Of the 144 patients in the development and validation datasets, only 7 (5%) who began treatment with DIBH required an unplanned conversion to free-breathing; for modelling purposes, these patients were classified as DIBH-ineligible. The assessment method was consistent across all socio-demographic groups to ensure fairness and accuracy. All analyses were performed using Python v3.11.[24] Decision curve analysis was performed in R v4.4.1 using the *rmda* package (https://github.com/mdbrown/rmda). All model development and evaluation was undertaken on a MacBook Pro (Apple M2, 8-core SoC, 10-core GPU, 24GB RAM). This report was prepared in accordance with the TRIPOD+AI guidelines (checklist in supplemental materials).[25]

**B) Model Development**

The development dataset (January 2023 to December 2023)(*n* = 202) was split into training and test sets in a 70:30 ratio, after stratifying for the outcome variable to ensure consistent distribution of events. Each algorithm's objective function was defined to maximize the 10-fold cross-validated Area Under the Curve (AUC) on the training set. The following supervised learning algorithms were evaluated for their efficacy in predicting the outcome: Logistic Regression (LR)[26]; Random Forest (RF)[27]: Gradient Boosting[28]; LightGBM (Light Gradient Boosting Machine)[29]; XGBoost (Extreme Gradient Boosting)[30]; CatBoost (Categorical Boosting)[31]; K-Nearest Neighbors (KNN)[32]; Support Vector Machines[33]; Naive Bayes[34]. Additional details are provided in supplemental materials.

For each algorithm, the model-building steps (figure 1) were as follows:

1. Hyperparameter optimization

Hyperparameter optimization (HPO) was conducted using *Optuna*, which allowed dynamic search space construction and efficient parameter tuning via the Tree-structured Parzen Estimator (TPE) algorithm.[35] Five separate HPO runs with different random seeds were performed to ensure robust optimization. The parameter sets from all five runs were used for further analysis.

1. Bootstrapping

The training set was resampled with replacement to generate 1000 unique bootstrap samples, while the test set was kept separate for evaluation. For each algorithm, the five HPO sets were used to train models on these 1000 bootstrap samples. This resulted in 5000 models per algorithm (1000 samples per HPO set). Additionally, a calibrated version of each model was created using the *CalibratedClassifierCV*. Both calibrated and uncalibrated models were evaluated during bootstrapping, with AUC and recall (for non-DIBH) serving as primary metrics. Each HPO set was further assessed by computing the average AUC, F1-score, precision, recall, and accuracy across 2000 models (1000 calibrated and 1000 uncalibrated). Classification thresholds were optimized using Youden's J-statistic. The best-performing HPO set was selected to create the ensemble models.

1. Ensemble model creation

The top 10 performing models from the selected HPO run (comprising 2000 models: 1000 calibrated and 1000 uncalibrated) were used to construct the ensemble model. The final ensemble prediction was determined by majority voting. In cases of a tie, the model defaulted to predicting DIBH eligibility. For each machine learning algorithm, three models were ultimately produced: an ensemble calibrated model, an ensemble uncalibrated model, and the single best uncalibrated model. These models were evaluated further to determine their suitability for predicting DIBH ineligibility in clinical practice.

**C) Model validation, evaluation, and decision curve analysis**

The performance of the final calibrated ensemble model, uncalibrated ensemble model, and the single best uncalibrated model for each algorithm was tested on the temporal validation dataset (*n* = 47). Models with high AUC were preferred in order to maximize discrimination. A decision curve analysis was performed to first determine the best model from the top 3 models in each group (calibrated ensemble models, uncalibrated ensemble models, and the single best uncalibrated models) and then the best model overall.[36]

After discussion amongst the investigators (KSC, AAM, IA, RB) it was decided that any threshold probability above which even a single patient is misclassified as non-DIBH was unacceptable (as it would result in a higher heart Dmean). From a statistical perspective, this would require a model with near-perfect discrimination.

This prompted the adoption of a more conservative approach: to set a threshold probability below which all patients would be labelled as DIBH. This decision acknowledged that some patients ineligible for DIBH might be misclassified as DIBH, but it ensures that no true DIBH patients are missed. This threshold probability was set at 0.2 (corresponding to a non-DIBH:DIBH ratio of 1:4) and served as the basis for selecting an overall best performing model.

Feature importance was quantified using a permutation-based approach. In brief, each feature was shuffled in isolation while keeping others constant, and the decline in performance was measured, to determine dependence on that feature. This model-agnostic approach was adopted to minimise bias from tree-based split frequencies and provided a reliable estimate of feature importance.

**D) Clinical impact study design**

A prospectively collected dataset (April 2024 to August 2024)(*n* = 64) was used to assess the best performing model’s clinical impact. First, we identified the threshold probability at which no more than one DIBH-ineligible patient was misclassified as DIBH-eligible. Next, the model was finalized and provided a binary output after assessment parameters for day 1 were entered: if predicted probability was less than or equal to the threshold, the model recommended 'consider CT simulation in DIBH'; if predicted probability was greater than the threshold, the model recommended 'suggest continuing DIBH assessment’. The model's output was available only to the principal investigators (KSC, AAM, IA).

The study was designed to evaluate how the model's predictions aligned with the team's decision, without influencing the decision itself. We evaluated alignment between the model’s prediction and the clinical team’s decision by comparing the final binary output (“consider CT simulation in DIBH” or “suggest continuing DIBH assessment”) with the standard clinical classification of each patient as DIBH-eligible or DIBH-ineligible. A prediction was considered correct if the model’s output matched the team’s designation, while any discrepancy was deemed a misclassification. Specifically, the clinical impact study measured how many patients could forgo assessment on 2nd and 3rd days without increasing the risk of misclassifying DIBH-ineligible patients.

**Results**

Descriptive demographics and measurements extracted from 1st day of assessment are shown in supplemental materials. This study evaluated multiple supervised learning algorithms and model development strategies for predicting DIBH ineligibility (summarised in Figure 1). The performance of the calibrated ensemble, uncalibrated ensemble, and single best models for each algorithm was compared based on the Area under the curve (AUC) on a temporal validation dataset. The top three best performing models from each strategy were then compared using a Decision Curve Analysis (DCA) at an *a priori* threshold of 0.2 (supplemental figure 2).

Each modelling strategy provided different optimal models. For both the calibrated and uncalibrated ensemble approaches, Gradient Boosting achieved the highest net benefit, with AUC (95% CI)/Recall of 0.790 (0.654–0.922)/0.526 and 0.803 (0.686–0.941)/0.526, respectively. Using the single best uncalibrated strategy, Extreme Gradient Boosting and LightGBM showed equivalent net benefit, with AUC (95% CI)/Recall values of 0.794 (0.664–0.918)/0.737 and 0.792 (0.657–0.916)/0.842, respectively.

On comparing the best models from all modelling strategies using a DCA (at 0.2 threshold), the uncalibrated Gradient Boosting ensemble model demonstrated the highest net benefit (figure 2A). The density curve of this model's predicted probability for DIBH and non-DIBH was also compared with the other models (figure 2B) and misclassification of a single non-DIBH patient as DIBH occurred at the predicted probability of 0.35 (corresponding to a non-DIBH:DIBH ratio of 1:14). The model was finalized at this stage and a clinical impact study was performed.

The importance of each feature in predicting DIBH ineligibility was analysed for the final uncalibrated Gradient Boosting ensemble model (figure 3). The top-ranking features were average breath-hold duration and lower level of amplitude. These were followed closely by average breath-hold amplitude and consistency of breath hold. This demonstrates the model’s reliance on length and consistency of breath-hold, as well as average breath-hold amplitude. These characteristics align with the clinical instinct to select patients who can reproducibly sustain a longer, deep inspiration breath-hold.

The importance of lower amplitude suggests that patients’ worst attempt strongly influences the model’s output, similar to the inclination to deem patients with low breath-hold amplitudes ineligible for treatment in DIBH. Age also contributed, consistent with prior observations of diminished pulmonary function in older patients. Interestingly, upper amplitude (i.e., a single “best” breath-hold) was less critical, showing that consistent breath-hold across attempts mattered more than a single maximal amplitude breath-hold.

The clinical impact study showed that the chosen threshold of the model resulted in equivalent performance, resulting in a non-DIBH:DIBH ratio of 1:12 (figure 4). By avoiding assessment on 2nd and 3rd days, 13 patients out of 64 in the clinical impact study could have proceeded directly to simulation in DIBH, which would have reduced workload by 20% (13/64 patients).

**Discussion**

This study reports the development, temporal validation, and clinical impact analysis of a machine learning model to aid in the prediction of Deep Inspiration Breath Hold (DIBH) ineligibility in patients with left-sided breast cancer, using clinical variables and DIBH waveform data extracted from the first day of assessment. Its development was motivated by the need for a decision aid to broaden DIBH implementation. Strengths of this study are the size of the training dataset derived from an LMIC institute, incorporation of DIBH assessment parameters, a comprehensive evaluation of a wide array of machine learning algorithms, and diverse model development strategies.

Unsuccessful DIBH RT can be resource-intensive when attempts are made to obtain optimal breathing. This will impact busy, resource-constrained departments. Our model accurately reduced the need for additional assessments for up to 20% of patients, as demonstrated in the clinical impact study. The unique challenges in LMICs comprise outdated radiotherapy infrastructure, and hence limitations on modern techniques. High-volume, modern LMIC institutions focus on benefitting a maximal aliquot of patients within their resources. Given abundant resources and lower patient volumes, a ‘default DIBH irrespective of breath-hold quality’ approach can be adopted, albeit with uncertainties in treatment delivery dosimetry.[37] Ultra-hypofractionated RT (UHFx) also promises greater throughput, though it is currently restricted to non-nodal breast/chest wall RT, which means that widespread application in LMICs (where nodal positivity at diagnosis is higher) may remain limited.[38,39] In all cases in LMIC institutions, time minimisation is necessary to maximise throughput and resource use.

To the best of our knowledge, the only precedent to this analysis is our previously reported logistic mixed model.[21] Remarkably, the feature importance of our machine learning model closely mirrors significant variables found in our logistic mixed model. However, this model’s utility is enhanced by requiring only the parameters from the 1st day of assessment, while the logistic mixed model requires data from all three days. Data derived from DIBH waveforms has also been used to quantify reduction in MHD and supports our assertion that effectively quantifying waveform can guide important clinical decisions.[40,41] Since the basis of both models is waveform data, our research group is actively investigating a deeper integration of our DIBH ineligibility model with an MHD reduction model. Preliminary research also demonstrates the potential of using a plain chest radiograph to predict MHD reduction and DIBH ineligibility.[20,42]

Exploring alternative end-points (cost savings, workflow reduction metrics) to assess the model’s performance is an avenue for future research, especially in LMICs where radiotherapy demand exceeds availability. An assessment of patient comfort and satisfaction with the DIBH procedure also needs to be addressed, and finally, we acknowledge the inability to provide per-fraction on-couch time for patients who underwent treatment in DIBH because the process of data extraction from our record-and-verify system is ongoing.

In this study, the model’s recommendations were visible only to the principal investigators after integration with our workflow management system, to avoid influencing clinical decisions. Since then, the model’s recommendations have been deployed in a live clinical environment under a separate clinical study, results from which will be available soon. We observed initial hesitancy among our staff in trusting the model’s recommendation, which was addressed by explaining the importance of maintaining independent clinical judgement. Moving forward, its single-institution design and relatively modest cohort size.

We acknowledge that inadvertently, demographic, clinical and socio-economic biases could have been introduced, potentially limiting model generalizability. The inter- and intra-observer variability in interpreting waveform data also needs to be acknowledged, for which we will provide an online, interactive demonstration via a non-profit organization (MedPy Foundation Labs). All these factors individually and taken together may have influenced our model’s performance. This will be addressed by adopting a federated learning framework, where other institutions using our model will not be required to share raw patient data, just the updated model weights will be used.

In conclusion, our study displays the potential of ML to improve selection for DIBH radiotherapy in patients with left-sided breast cancer by integrating clinical and waveform assessment data. While limitations exist, we are optimistic that the radiation oncology research community will build upon our results and develop more solutions specifically addressing challenges faced in LMICs.

**References**

1 Aznar MC, Carrasco De Fez P, Corradini S, *et al.* ESTRO-ACROP guideline: Recommendations on implementation of breath-hold techniques in radiotherapy. *Radiother Oncol*. 2023;185:109734. doi: 10.1016/j.radonc.2023.109734

2 Anastasi G, Bertholet J, Poulsen P, *et al.* Patterns of practice for adaptive and real-time radiation therapy (POP-ART RT) part I: Intra-fraction breathing motion management. *Radiother Oncol*. 2020;153:79–87. doi: 10.1016/j.radonc.2020.06.018

3 Lai J, Hu S, Luo Y, *et al.* Meta-analysis of deep inspiration breath hold (DIBH) versus free breathing (FB) in postoperative radiotherapy for left-side breast cancer. *Breast Cancer*. 2020;27:299–307. doi: 10.1007/s12282-019-01023-9

4 Darby SC, Ewertz M, McGale P, *et al.* Risk of Ischemic Heart Disease in Women after Radiotherapy for Breast Cancer. *N Engl J Med*. 2013;368:987–98. doi: 10.1056/NEJMoa1209825

5 Bartlett FR, Donovan EM, McNair HA, *et al.* The UK HeartSpare Study (Stage II): Multicentre Evaluation of a Voluntary Breath-hold Technique in Patients Receiving Breast Radiotherapy. *Clin Oncol*. 2017;29:e51–6. doi: 10.1016/j.clon.2016.11.005

6 Chatterjee S, Chakraborty S, Moses A, *et al.* Resource requirements and reduction in cardiac mortality from deep inspiration breath hold (DIBH) radiation therapy for left sided breast cancer patients: A prospective service development analysis. *Pract Radiat Oncol*. 2018;8:382–7. doi: 10.1016/j.prro.2018.03.007

7 Kron T, Bressel M, Lonski P, *et al.* TROG 14.04: Multicentre Study of Feasibility and Impact on Anxiety of DIBH in Breast Cancer Patients. *Clin Oncol*. 2022;34:e410–9. doi: 10.1016/j.clon.2022.05.020

8 Lin C-H, Lin L-C, Que J, *et al.* A seven-year experience of using moderate deep inspiration breath-hold for patients with early-stage breast cancer and dosimetric comparison. *Medicine (Baltimore)*. 2019;98:e15510. doi: 10.1097/MD.0000000000015510

9 Tanguturi SK, Lyatskaya Y, Chen Y, *et al.* Prospective assessment of deep inspiration breath-hold using 3-dimensional surface tracking for irradiation of left-sided breast cancer. *Pract Radiat Oncol*. 2015;5:358–65. doi: 10.1016/j.prro.2015.06.002

10 Locke I, Drinkwater K. Implementation of Royal College of Radiologists Consensus Statements and National Institute for Health and Care Excellence Guidance: Breast Radiotherapy Practice in the UK. *Clin Oncol*. 2021;33:419–26. doi: 10.1016/j.clon.2021.01.012

11 Vikström J, Hjelstuen MHB, Mjaaland I, *et al.* Cardiac and pulmonary dose reduction for tangentially irradiated breast cancer, utilizing deep inspiration breath-hold with audio-visual guidance, without compromising target coverage. *Acta Oncol*. 2011;50:42–50. doi: 10.3109/0284186X.2010.512923

12 Kim A, Kalet AM, Cao N, *et al.* Effects of Preparatory Coaching and Home Practice for Deep Inspiration Breath Hold on Cardiac Dose for Left Breast Radiation Therapy. *Clin Oncol*. 2018;30:571–7. doi: 10.1016/j.clon.2018.04.009

13 Li H, Zhang F, Qiu J. Benefits of Regular Breath-Holding Training for DIBH Treatment. *Int J Radiat Oncol*. 2024;120:e152–3. doi: 10.1016/j.ijrobp.2024.07.2120

14 Bartlett FR, Colgan RM, Carr K, *et al.* The UK HeartSpare Study: Randomised evaluation of voluntary deep-inspiratory breath-hold in women undergoing breast radiotherapy. *Radiother Oncol*. 2013;108:242–7. doi: 10.1016/j.radonc.2013.04.021

15 Drost L, Yee C, Lam H, *et al.* A Systematic Review of Heart Dose in Breast Radiotherapy. *Clin Breast Cancer*. 2018;18:e819–24. doi: 10.1016/j.clbc.2018.05.010

16 Chufal KS, Ahmad I, Miller AA, *et al.* A practical and practicable framework for implementing cardiac-sparing radiotherapy techniques in breast cancer. *Cancer Res Stat Treat*. 2023;6:492–5. doi: 10.4103/crst.crst\_33\_23

17 van den Bogaard VAB, Ta BDP, van der Schaaf A, *et al.* Validation and Modification of a Prediction Model for Acute Cardiac Events in Patients With Breast Cancer Treated With Radiotherapy Based on Three-Dimensional Dose Distributions to Cardiac Substructures. *J Clin Oncol*. 2017;35:1171–8. doi: 10.1200/JCO.2016.69.8480

18 Tanna N, McLauchlan R, Karis S, *et al.* Assessment of Upfront Selection Criteria to Prioritise Patients for Breath-hold Left-sided Breast Radiotherapy. *Clin Oncol*. 2017;29:356–61. doi: 10.1016/j.clon.2017.01.004

19 Zhou Y, Xu J, Xu F, *et al.* Selection criteria and method for deep inspiration breath-hold in patients with left breast cancer undergoing PMRT/IMRT. *Clin Transl Radiat Oncol*. 2024;48:100812. doi: 10.1016/j.ctro.2024.100812

20 Koide Y, Aoyama T, Shimizu H, *et al.* Development of deep learning chest X-ray model for cardiac dose prediction in left-sided breast cancer radiotherapy. *Sci Rep*. 2022;12:13706. doi: 10.1038/s41598-022-16583-8

21 Ahmad I, Chufal KS, Miller AA, *et al.* Identification of variables and development of a prediction model for DIBH eligibility in left-sided breast cancer radiotherapy: a prospective cohort study with temporal validation. *Radiat Oncol*. 2024;19:115. doi: 10.1186/s13014-024-02512-8

22 Zubizarreta EH, Fidarova E, Healy B, *et al.* Need for Radiotherapy in Low and Middle Income Countries – The Silent Crisis Continues. *Clin Oncol*. 2015;27:107–14. doi: 10.1016/j.clon.2014.10.006

23 Yusuf S, Rangarajan S, Teo K, *et al.* Cardiovascular Risk and Events in 17 Low-, Middle-, and High-Income Countries. *N Engl J Med*. 2014;371:818–27. doi: 10.1056/NEJMoa1311890

24 Pedregosa F, Varoquaux G, Gramfort A, *et al.* Scikit-learn: Machine Learning in Python. *J Mach Learn Res*. 2011;12:2825–30.

25 Collins GS, Moons KGM, Dhiman P, *et al.* TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*. 2024;e078378. doi: 10.1136/bmj-2023-078378

26 Walker SH, Duncan DB. Estimation of the Probability of an Event as a Function of Several Independent Variables. *Biometrika*. 1967;54:167. doi: 10.2307/2333860

27 Tin Kam Ho. The random subspace method for constructing decision forests. *IEEE Trans Pattern Anal Mach Intell*. 1998;20:832–44. doi: 10.1109/34.709601

28 Friedman JH. Greedy function approximation: A gradient boosting machine. *Ann Stat*. 2001;29. doi: 10.1214/aos/1013203451

29 Ke G, Meng Q, Finley T, *et al.* LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In: Guyon I, Luxburg UV, Bengio S, *et al.*, eds. *Advances in Neural Information Processing Systems*. Curran Associates, Inc. 2017.

30 Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco California USA: ACM 2016:785–94.

31 Prokhorenkova L, Gusev G, Vorobev A, *et al.* CatBoost: unbiased boosting with categorical features. 2017.

32 Cover T, Hart P. Nearest neighbor pattern classification. *IEEE Trans Inf Theory*. 1967;13:21–7. doi: 10.1109/TIT.1967.1053964

33 Cortes C, Vapnik V. Support-vector networks. *Mach Learn*. 1995;20:273–97. doi: 10.1007/BF00994018

34 Langley P, Iba W, Thompson K. An Analysis of Bayesian Classifiers. *AAAI Conference on Artificial Intelligence*. 1992.

35 Akiba T, Sano S, Yanase T, *et al.* Optuna: A Next-generation Hyperparameter Optimization Framework. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. Anchorage AK USA: ACM 2019:2623–31.

36 Vickers AJ, Elkin EB. Decision Curve Analysis: A Novel Method for Evaluating Prediction Models. *Med Decis Making*. 2006;26:565–74. doi: 10.1177/0272989X06295361

37 Kügele M, Edvardsson A, Berg L, *et al.* Dosimetric effects of intrafractional isocenter variation during deep inspiration breath‐hold for breast cancer patients using surface‐guided radiotherapy. *J Appl Clin Med Phys*. 2018;19:25–38. doi: 10.1002/acm2.12214

38 Meattini I, Becherini C, Boersma L, *et al.* European Society for Radiotherapy and Oncology Advisory Committee in Radiation Oncology Practice consensus recommendations on patient selection and dose and fractionation for external beam radiotherapy in early breast cancer. *Lancet Oncol*. 2022;23:e21–31. doi: 10.1016/S1470-2045(21)00539-8

39 Bhattacharyya GS, Doval DC, Desai CJ, *et al.* Overview of Breast Cancer and Implications of Overtreatment of Early-Stage Breast Cancer: An Indian Perspective. *JCO Glob Oncol*. 2020;789–98. doi: 10.1200/GO.20.00033

40 Ledsom D, Reilly AJ, Probst H. Assessment of deep inspiration breath hold (DIBH) amplitude and reduction in cardiac dose in left breast cancer patients. *Radiography*. 2018;24:98–103. doi: 10.1016/j.radi.2017.11.005

41 Vendrame A, Cappelletto C, Chiovati P, *et al.* Artificial Intelligence-Based Patient Selection for Deep Inspiration Breath-Hold Breast Radiotherapy from Respiratory Signals. *Appl Sci*. 2023;13:4962. doi: 10.3390/app13084962

42 Chufal KS, Ahmad I, Sharief MI, *et al.* PD-0730 Convolutional Neural Network to predict Deep Inspiration Breath Hold eligibility using Chest X-Ray. *Radiother Oncol*. 2021;161:S560–1. doi: 10.1016/S0167-8140(21)07009-2

**Figure Captions**

**Figure 1: Illustration of the model development process, including hyperparameter optimization (HPO), bootstrapping, and ensemble model creation.**



The dataset was split into training (n=141) and test (n=61) sets in a 70:30 ratio. Five HPO runs were conducted for each algorithm. Each HPO run produced five parameter sets that were used to train models on 1000 bootstrapped samples. For each algorithm, both calibrated and uncalibrated models were evaluated based on AUC and recall. The top 10 models from the best-performing HPO set were selected for ensemble model creation. Three final models were produced for each algorithm: a calibrated ensemble model, an uncalibrated ensemble model, and the single best uncalibrated model. These models were tested on the validation dataset (n=47).

**Figure 2: Decision curve analysis and Predicted probability density**



(A) Decision curve analysis (DCA) comparing the net benefit of the top models from three different model-building strategies: calibrated ensemble, uncalibrated ensemble, and single best uncalibrated models. The uncalibrated Gradient Boosting ensemble model (green) demonstrated the highest net benefit at the 0.2 threshold probability, outperforming the other models across all strategies. (B) Density curves of the predicted probability distributions for DIBH and non-DIBH for the top models from each strategy. The uncalibrated Gradient Boosting model (solid red) showed a clear separation between DIBH and non-DIBH patients till the chosen threshold of 0.35.

**Figure 3: Feature importance of the final uncalibrated Gradient Boosting ensemble model**



The relative importance of each feature in predicting DIBH ineligibility across 10 models within the uncalibrated Gradient Boosting ensemble is shown. The top-ranking features were average breath-hold duration [Breath Hold (Average duration)], lower amplitude [Amplitude (LL)], average breath-hold amplitude [Amplitude (Average)] and consistent breath-hold [Breath hold (Consistency)]. Patient-related factors, such as age, BMI, surgery type, comorbidities, and chemotherapy, showed lower importance compared to DIBH assessment parameters, except upper amplitude [Amplitude (UL)]. For a detailed description of the predictors, please see [21] and supplemental materials.

**Figure 4: Comparison of predicted probabilities in temporal validation and clinical impact datasets**



The plot shows the predicted probabilities of Deep Inspiration Breath Hold (DIBH) and non-DIBH patients in the temporal validation dataset (left) and clinical impact dataset (right). Blue circles represent DIBH patients, and red circles represent non-DIBH patients. The purple dashed line indicates the chosen threshold (predicted probability = 0.35), where misclassification of a single non-DIBH patient as DIBH occurred, corresponding to a non-DIBH ratio of 1:14. At this threshold, its simulated results from the clinical impact dataset demonstrated a similar non-DIBH ratio of 1:12, which could have reduced the need for additional DIBH assessments by 20% (13/64 patients).