Supplementary Material: Breast cancer risk prediction from longitudinal mammography reports in a real-world health system

Higor S. Monteiro,[1] José A. Shiomi da Cruz,[2,3] César L. C. Mattos,[4] Iago C. Chaves,[4] Javam C. Machado,[4] Hernán A. Makse,[5] Lucas C. Parra,[6] and José S. Andrade Jr.[1]

[1]*Departamento de Física, Universidade Federal do Ceará, Campus do Pici, 60451-970 Fortaleza, Ceará, Brazil*

[2]*Hapvida Assistência Médica S. A., 60140-061, Fortaleza, Ceará, Brazil*

[3]*Departamento de Urologia, Escola Paulista de Medicina, Universidade Federal de São Paulo, 04023-900 São Paulo, Brazil*

[4]*Departmento de Computação, Universidade Federal do Ceará, Campus do Pici, 60451-970 Fortaleza, Ceará, Brazil*

[5]*Levich Institute and Physics Department, City College of New York, New York, NY 10031, US*

[6]*Department of Biomedical Engineering, The City College of New York, New York, NY 10031, US*

CONTENTS

S1. Further context 3

S2. Extended Methods 4

S2.1. Hyperparameter Tuning 4

S2.2. TF-IDF Text Representation 5

S2.3. Lookback Window 6

S2.4. Only Index Mammogram Ablation Analysis 7

S2.5. Terms Related to Prior Breast Cancer Interventions 7

S2.6. Text-free Ablation Analysis 9

S2.7. Aggregation Function 9

S2.8. Biopsy Report Classification 10

S3. Sensitivity Analysis 11

S3.1. Effect of the BI-RADS 5 Confirmation Window 11

S3.2. Effect of the Post-Index Exclusion Window 12

S3.3. Performance Under Alternative BI-RADS Groupings 12

S3.4. Risk Heterogeneity Within BI-RADS 0 13

S4. Extended Explainability Analysis 14

References 16

S1. FURTHER CONTEXT

Figure S1 shows the distribution of time intervals between consecutive mammography examinations in the cohort. The distribution peaks around 12 months, consistent with annual screening recommendations, with a long tail extending beyond 24 months reflecting the heterogeneous screening adherence patterns observed in a real-world health system. Figure S2 shows the cancer detection rate by BI-RADS category as a function of prediction horizon, demonstrating that although BI-RADS 3 carries the highest per-category detection rate, BI-RADS 1 and 2 together account for the majority of cancer cases at the 5-year horizon (838 and 959 cases respectively, compared to 671 for BI-RADS 3).

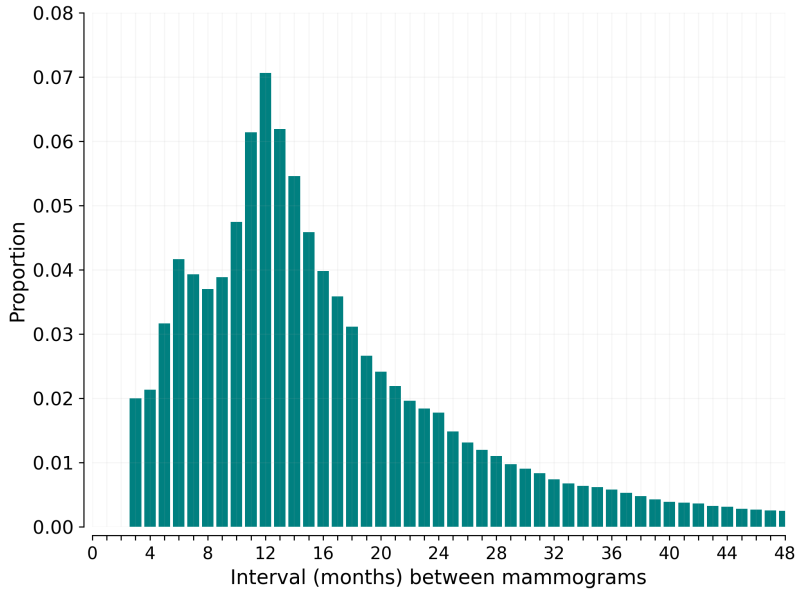

FIG. S1. Distribution of time intervals between consecutive mammography examinations in the cohort. The distribution peaks around 12 months, consistent with annual screening recommendations, with a long tail reflecting heterogeneous screening adherence patterns in a real-world health system.

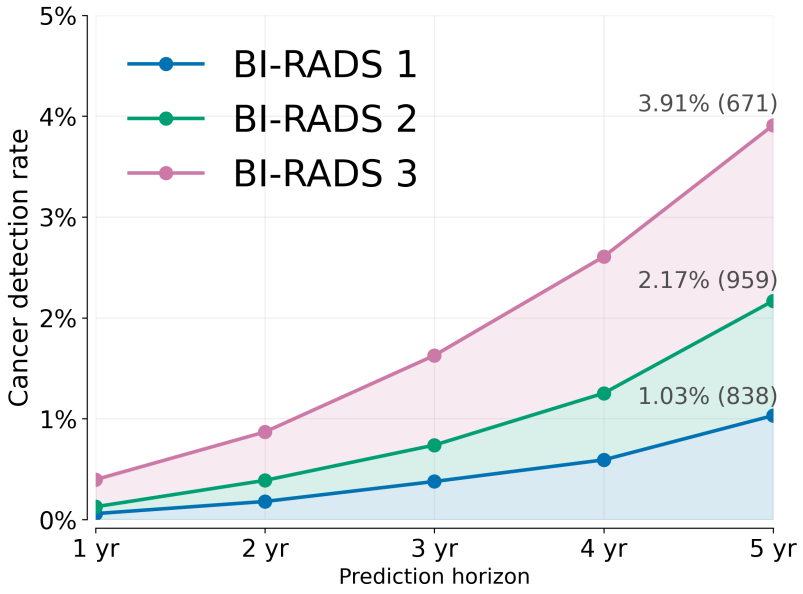

FIG. S2. Cancer detection rate by BI-RADS category as a function of prediction horizon. Although BI-RADS 3 carries the highest per-category detection rate, BI-RADS 1 and 2 together account for the majority of cancer cases at the 5-year horizon (838 and 959 cases respectively, compared to 671 for BI-RADS 3), highlighting that cancer burden is not concentrated in the highest BI-RADS category.

S2. EXTENDED METHODS

S2.1. Hyperparameter Tuning

Hyperparameter tuning was performed using the validation set via the Optuna framework [1], a hyperparameter optimization library based on a tree-structured Parzen estimator (TPE) sampler. The following hyperparameters were optimized: learning rate, weight decay, number of dense layers and their respective sizes, dropout rate, optimizer (Adam [2] or SGD [3]), and activation function (GeLU [4], ReLU [5], and Mish [6]). The final model for each prediction horizon was selected based on the best AUROC on the validation set. Optimization based on the area under the precision-recall curve (AUPRC) was also evaluated as an alternative objective, but was found to exhibit substantial fluctuation across trials, likely due to the low cancer prevalence within the validation set, and was therefore not adopted as the primary optimization criterion. Final hyperparameter values are reported in Table S1.

S2.2. TF-IDF Text Representation

Mammography report texts were preprocessed by removing punctuation and converting all characters to lowercase. A TF-IDF vectorizer was fitted on a representative sample of reports spanning all BI-RADS categories and all years of the study period, with a maximum vocabulary size of 5,000 terms, a minimum document frequency of 5, a maximum document frequency of 99.5%, and a unigram and bigram token range. Each report was represented as a sparse TF-IDF vector of dimensionality 5,000. For patients with multiple past reports within the lookback window, each report was independently encoded and the resulting embeddings were passed to the aggregation function together with their corresponding time intervals.

To ensure the robustness of the TF-IDF representation, we verified that IDF weights and vocabulary composition are stable with respect to the choice of documents used for fitting. IDF weights fitted on the training set only were compared against those fitted on the full corpus, yielding a Pearson correlation of 0.985, confirming that the feature space is not meaningfully affected by this choice and that no outcome-related leakage was introduced in the vocabulary construction process (Fig. S3).

TABLE S1. Final hyperparameter values for each prediction horizon, selected based on validation set AUROC.

Hyperparameter	**1-year**	**2-year**	**3-year**	**4-year**	**5-year**
Learning rate	6.07×10^{-4}	2.10×10^{-5}	1.19×10^{-5}	1.95×10^{-3}	1.64×10^{-5}
Weight decay	1.14×10^{-5}	1.81×10^{-5}	4.82×10^{-5}	7.76×10^{-5}	1.60×10^{-5}
Number of dense layers	3	4	5	3	2
Trainable parameters	150,945	2,891,297	3,940,897	343,233	5,286,145
Dropout rate	0.254	0.188	0.142	0.067	0.566
Optimizer	Adam	SGD	Adam	SGD	SGD
Activation function	ReLU	ReLU	ReLU	ReLU	ReLU

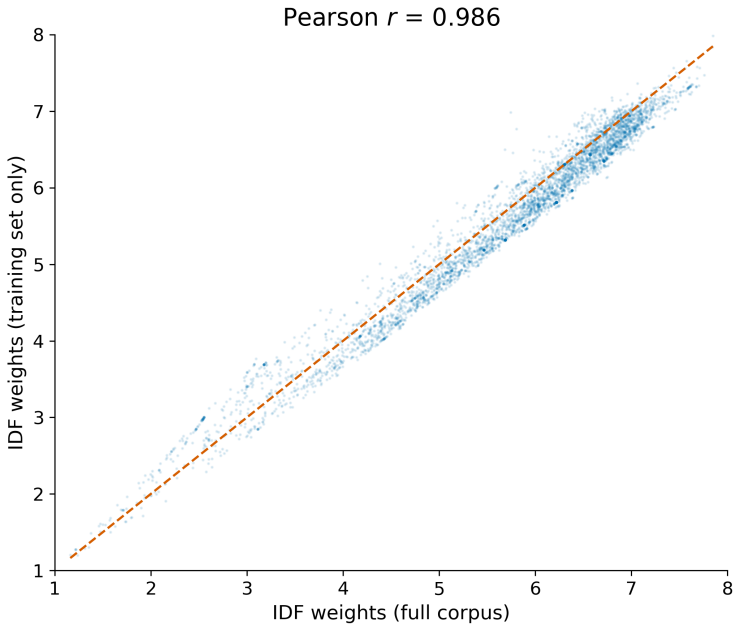

FIG. S3. Scatter plot of IDF weights fitted on the full corpus versus the training set only, across all common terms. The dashed line represents perfect agreement. Pearson correlation = 0.985, confirming the stability of the TF-IDF feature space with respect to the documents used for fitting.

TABLE S2. Model discrimination performance across different lookback window sizes for the 5-year prediction horizon on the held-out test set. The selected lookback window used in the primary analysis is indicated in bold.

Lookback window (years)	**AUROC (95% CI)**	**Concordance index (95% CI)**
2	0.852 (0.841–0.863)	0.848 (0.840–0.862)
3	**0.863 (0.853–0.872)**	**0.859 (0.850–0.867)**
5	0.862 (0.851–0.869)	0.851 (0.848–0.860)
10	0.864 (0.851–0.870)	0.856 (0.849–0.863)

S2.3. Lookback Window

The lookback window defines the maximum time interval before the index exam within which past mammography reports are considered as model input. A lookback window of 3 years was used in the primary analysis. The sensitivity of model performance to different lookback window sizes was assessed and results are reported in Table S2.

S2.4. Only Index Mammogram Ablation Analysis

To evaluate the contribution of the longitudinal aspect of the mammography reports, we trained a baseline model using only the index mammogram report as input. The Table S3 reports the performance for the full model and the index-only model for the 1-year and 5-year prediction horizons in the held-out test set.

TABLE S3. Discrimination performance for the full model and the index-only model for the 1-year and 5-year prediction horizons on the held-out test set.

Model	**1-year**	**5-year**
	AUROC (95% CI)	**AUROC (95% CI)**
Full model	0.849 (0.833–0.867)	0.858 (0.847–0.866)
Index mammogram only	0.802 (0.780–0.824)	0.855 (0.847–0.864)

S2.5. Terms Related to Prior Breast Cancer Interventions

The explainability analysis identified surgical history terms in the reports as strong contributors to the positive predictions, raising the concern that these terms may proxy a prior breast cancer history not consistently captured in the structured electronic health records. To assess whether such terms constitute a possible shortcut, we identified the examinations containing at least one term related to a possible prior breast cancer using a list of terms: *mastectomia* (mastectomy), *mastectomia radical* (radical mastectomy), *quadrantectomia* (quadrantectomy), *cirurgia de mama* (breast surgery), *cirurgia mamária* (breast surgery), *tumorectomia* (tumorectomy/lumpectomy), *nodulectomia* (nodulectomy), *exérese* (excision), *prótese mamária* (breast implant), *implante mamário* (breast implant), *reconstrução mamária* (breast reconstruction), and *cicatriz cirúrgica* (surgical scar). Examinations containing at least one of these terms accounted for approximately 8% of all positive cases in the held-out test set for the 5-year prediction. The model was retrained after excluding all such examinations from the training and test sets, and performance was evaluated on the remaining held-out test set. ROC curves for the 5-year and 1-year prediction horizons are shown in Figure **??**, and their values along with confidence intervals

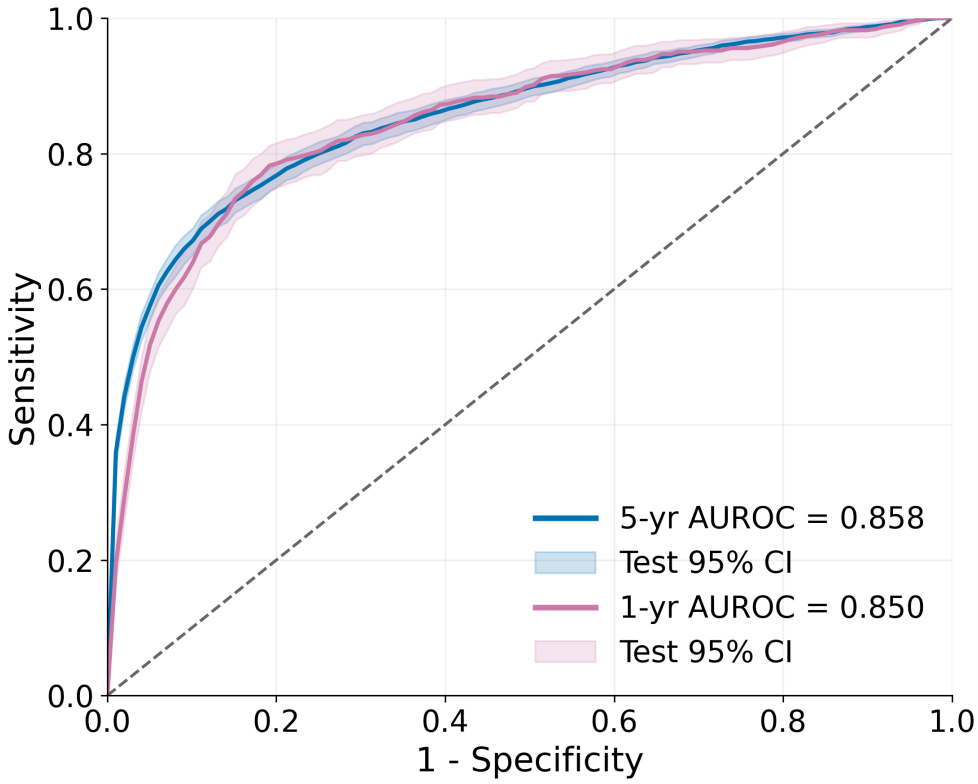

FIG. S4. ROC curves after retraining the model excluding the reports associated with previous breast cancer history.

are reported in Table S4.

TABLE S4. Model discrimination performance before and after excluding examinations containing terms related to prior breast cancer interventions, for the 1-year and 5-year prediction horizons on the held-out test set.

Model	**1-year AUROC (95% CI)**	**5-year AUROC (95% CI)**
Full model	0.844 (0.851–0.862)	0.863 (0.853–0.872)
Excluding prior intervention terms	0.849 (0.833–0.867)	0.858 (0.847–0.866)

S2.6. Text-free Ablation Analysis

To quantify the contribution of the reports' text to the model performance, we trained a text-free model using only structured clinical features extracted from the electronic health records, with the text embedding block replaced by a zero vector. Table S5 reports discrimination performance for the full model and the text-free model for the 1-year and 5-year prediction horizons on the held-out test set.

TABLE S5. Text-free ablation analysis for the 1-year and 5-year prediction horizons on the held-out test set.

Model	**1-year**	**5-year**
	AUROC (95% CI)	**AUROC (95% CI)**
Full model	0.844 (0.851–0.862)	0.863 (0.853–0.872)
Text-free model	0.755 (0.735–0.774)	0.791 (0.781– 0.803)

S2.7. Aggregation Function

The aggregation function summarizes the sequence of past mammography report embeddings and inter-examination time intervals into a single representative vector. Two aggregation strategies were evaluated: a simple mean of all past report embeddings, and a time-weighted mean where each embedding is weighted by an exponential decay function of the time interval between the corresponding past exam and the index exam. Formally, the weight assigned to a past exam at time t relative to the index exam at time t_0 is given by:

$$w(t) = \exp\left(-\lambda \cdot (t_0 - t)\right) \tag{1}$$

where $\lambda > 0$ controls the rate of decay. A higher value of λ causes the weights to decay more rapidly with time, meaning the model places greater emphasis on more recent examinations and discounts older ones more aggressively. Conversely, a lower value of λ results in a slower decay, giving more uniform weight to all past examinations regardless of how long ago they were performed. In the limit as $\lambda \to 0$, the time-weighted mean converges to the simple mean. The sensitivity of model performance to different values of λ was assessed and results

TABLE S6. Model discrimination performance across different aggregation strategies and decay parameter values for the 5-year prediction horizon on the held-out test set. The selected configuration used in the primary analysis is indicated in bold.

Aggregation strategy	**AUROC (95% CI)**	**Concordance index (95% CI)**
Simple mean	**0.863 (0.853–0.872)**	**0.859 (0.850–0.867)**
Time-weighted mean ($\lambda = 0.1$)	0.853 (0.841–0.865)	0.849 (0.843–0.862)
Time-weighted mean ($\lambda = 0.01$)	0.858 (0.845–0.863)	0.851 (0.844–0.865)
Time-weighted mean ($\lambda = 0.001$)	0.862 (0.855–0.872)	0.855 (0.843–0.865)

are reported in Table S6.

S2.8. Biopsy Report Classification

Biopsy reports were used to assign positive outcome labels to mammography examinations where no BI-RADS 5 or 6 classification was available within the prediction horizon. Prior to classification, all biopsy reports were deidentified, including the removal of pathologist names and any other personally identifiable information present in the report text. To classify each biopsy report as benign or malignant, we developed a prompt-based classification pipeline using large language models (LLMs). A ground-truth dataset of 1,000 biopsy reports was manually annotated by a physician and used to evaluate the classification performance of two LLM-based classifiers: the ChatGPT API (GPT-4) [7] and Qwen3:4b [8], an open-source model deployed locally via Ollama. Both models were prompted with a standardized zero-shot prompt instructing them to classify each report as benign or malignant based on the report text alone. Classification performance was assessed against the physician-annotated ground truth using accuracy, precision, recall, and F1-score, as reported in Table S7. Based on these results, the Qwen3:4b model was selected for the primary analysis given its strong performance and practical advantages in terms of cost, data privacy, and local deployment without reliance on external APIs.

TABLE S7. Performance of LLM-based biopsy report classifiers evaluated against a physician-annotated ground truth of 1,000 biopsy reports (786 benign, 214 malignant). Metrics are reported separately for benign (class 0) and malignant (class 1) classifications.

Classifier	**Class**	**Precision**	**Recall**	**F1-score**
ChatGPT API (GPT-4)	Benign	0.99	0.99	0.99
	Malignant	0.97	0.98	0.98
Qwen3:4b (Ollama)	Benign	0.99	0.99	0.99
	Malignant	0.97	0.98	0.98

S3. SENSITIVITY ANALYSIS

S3.1. Effect of the BI-RADS 5 Confirmation Window

In the primary analysis, a BI-RADS 5 classification was treated as a positive outcome only if no subsequent negative or benign mammogram was recorded within 6 months of the index exam, to avoid misclassifying cases where the BI-RADS 5 finding was subsequently reassessed as benign. To assess the sensitivity of this design choice, we tested alternative confirmation windows of 3, 9, and 12 months. As shown in Table S8, model performance was stable across all tested windows, supporting the robustness of the primary outcome definition.

TABLE S8. Model discrimination performance across different BI-RADS 5 confirmation windows for the 5-year prediction horizon on the held-out test set. The selected interval used in the primary analysis is indicated in bold.

Confirmation window (months)	**AUROC (95% CI)**	**C-index (95% CI)**
3	0.862 (0.854–0.872)	0.861 (0.856–0.872)
6	**0.863 (0.853–0.872)**	**0.859 (0.850–0.867)**
9	0.861 (0.852–0.872)	0.859 (0.850–0.866)
12	0.860 (0.850–0.871)	0.855 (0.849–0.864)

S3.2. Effect of the Post-Index Exclusion Window

To avoid including examinations where a malignant tumour was already present at the time of the index mammogram but not detected by the radiologist, we excluded all exams in which a confirmed malignancy was recorded within a defined interval after the index exam. In the primary analysis, this interval was set to 90 days. To assess the sensitivity of model performance to this design choice, we trained and evaluated the model using alternative exclusion intervals of 30, 60, and 120 days. As shown in Table S9, no significant differences in model performance were observed across these intervals, with the greatest impact arising from the shortest interval of 30 days, likely reflecting a small number of borderline cases near the boundary.

S3.3. Performance Under Alternative BI-RADS Groupings

In the primary analysis, the cohort was restricted to examinations classified as BI-RADS 1, 2, or 3. To assess the generalizability of the model to alternative cohort definitions, we also trained and evaluated the model on two additional groupings: a more restrictive cohort including only BI-RADS 1 and 2 examinations, and a more inclusive cohort incorporating BI-RADS 0 examinations alongside the primary categories. Results are reported in Table S10. The model achieved strong discrimination across all cohort definitions, with results comparable to or exceeding those reported in the literature for similar approaches.

TABLE S9. Model discrimination performance across different post-index exclusion windows for the 5-year prediction horizon on the held-out test set. The selected interval used in the primary analysis is indicated in bold.

Exclusion window (days)	**AUROC (95% CI)**	**C-index (95% CI)**
30	0.853 (0.846–0.863)	0.842 (0.832–0.850)
60	0.840 (0.831–.850)	0.840 (0.832–0.853)
90	**0.863 (0.853–0.872)**	**0.859 (0.850–0.867)**
120	0.851 (0.842–0.860)	0.839 (0.831–0.847)

TABLE S10. Model discrimination and operational performance across alternative BI-RADS cohort definitions for the 5-year prediction horizon on the held-out test set. The cohort definition used in the primary analysis is indicated in bold. Sensitivity, specificity, and PPV are reported at the top 3.43% risk threshold.

Metric	**1 and 2 only**	**1, 2 and 3**	**0, 1, 2 and 3**
AUROC (95% CI)	0.825 (0.814–0.835)	**0.863 (0.853–0.872)**	0.851 (0.844–0.860)
C-index (95% CI)	0.812 (0.809–0.820)	**0.859 (0.850–0.867)**	0.841 (0.833–0.848)
Sensitivity (95% CI)	40.7% (38.8%–42.6%)	**50.0% (48.4%–52.2%)**	41.4% (40.0%–43.0%)
Specificity (95% CI)	97.0% (97.0%–97.1%)	**97.3% (97.3%–97.4%)**	97.3% (97.2%–97.3%)
PPV (95% CI)	16.8% (15.7%–17.5%)	**25.1% (24.1%–26.3%)**	23.5% (22.7%-24.6%)

S3.4. Risk Heterogeneity Within BI-RADS 0

In the primary analysis, examinations classified as BI-RADS 0 were excluded from the cohort, as this category indicates an incomplete assessment requiring additional imaging, making immediate clinical management ambiguous. Nevertheless, to assess whether the model captures meaningful risk heterogeneity within this category, we applied the trained model to BI-RADS 0 examinations from the held-out test set and examined the distribution of predicted 5-year breast cancer risk scores. As shown in Figure S5, among true positive exams, BI-RADS 0 displays a broad and elevated risk score distribution in the high-risk range, comparable to that observed for BI-RADS 1, 2, and 3, suggesting that the model identifies clinically meaningful risk differences even among examinations with inconclusive radiological findings. Figure 2 in the main text further shows that across all exams, BI-RADS 0 is assigned substantially higher risk scores on average compared to BI-RADS 1, consistent with its more ambiguous clinical status. Together, these results are consistent with the risk heterogeneity observed within the primary BI-RADS categories reported in the main text, and further support the complementary value of the model beyond standard radiological classification.

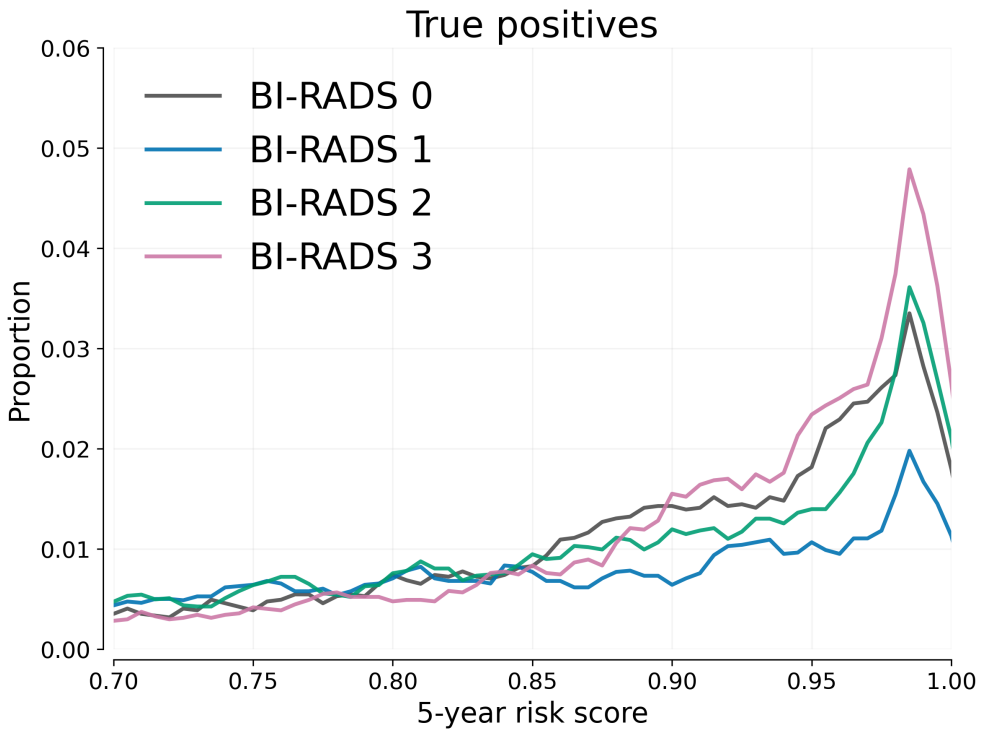

FIG. S5. Risk score distributions including BI-RADS 0. True positives across all BI-RADS categories. Smoothed risk score distributions in the high-risk range are shown for true positive exams stratified by BI-RADS category, including BI-RADS 0 alongside the primary analysis categories (1, 2, and 3). BI-RADS 0 exhibits a broad and elevated risk distribution comparable to the other categories, suggesting the model captures meaningful risk heterogeneity even among examinations with inconclusive radiological findings.

S4. EXTENDED EXPLAINABILITY ANALYSIS

Figure S6 and S7 shows the individual attribution scores for each structured feature across the four prediction outcome groups (true positives, false positives, true negatives, and false negatives). Attributions were computed using the integrated gradients method as described in the main text. Features are ranked by mean absolute attribution across the full test set. This analysis complements the pair-wise attribution heatmaps shown in the main text by providing a more granular view of the individual contribution of each structured risk factor to model predictions across outcome groups.

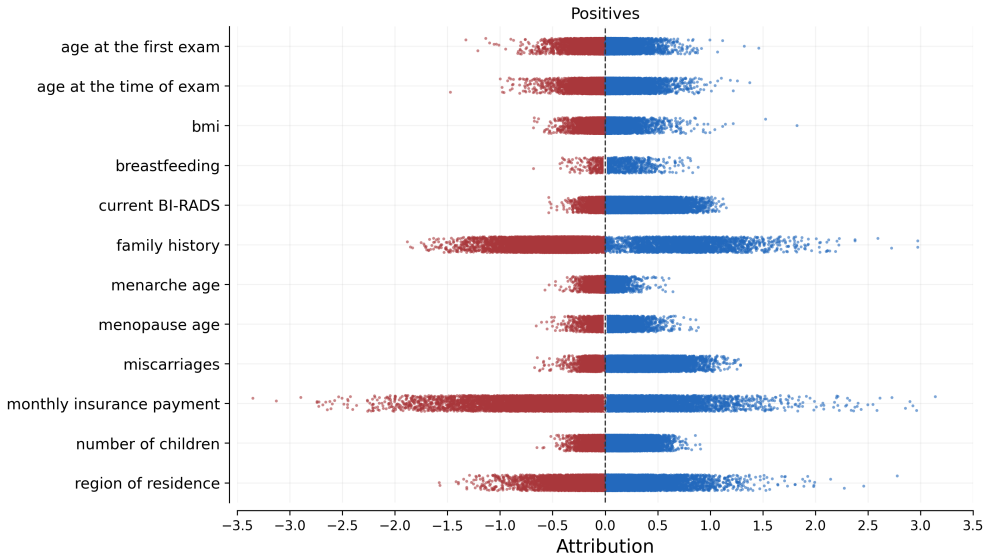

FIG. S6. Individual integrated gradients attribution scores for structured features in the group predicted as positive on the held-out test set. Each point represents one exam. The vertical dashed line indicates zero attribution.

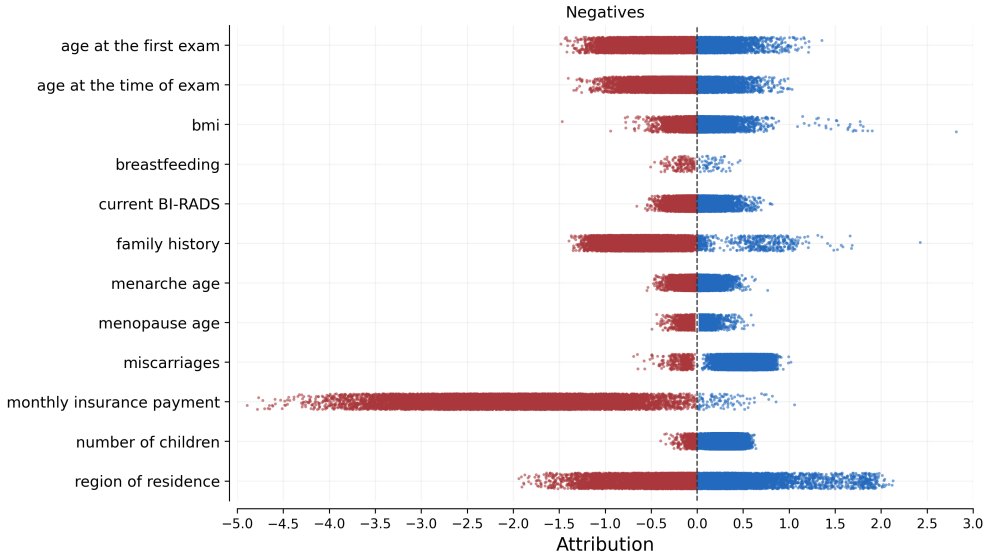

FIG. S7. Individual integrated gradients attribution scores for structured features in the group predicted as negative on the held-out test set. Each point represents one exam. The vertical dashed line indicates zero attribution.

[1] Takuya Akiba, Shotaro Sano, Toshihiro Yanase, Tetsuro Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2623–2631. ACM, 2019. doi:10.1145/3292500.3330701. URL `https://doi.org/10.1145/3292500.3330701`.

[2] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. URL `https://arxiv.org/abs/1412.6980`.

[3] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951. doi:10.1214/aoms/1177729586.

[4] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. URL `https://arxiv.org/abs/1606.08415`.

[5] Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted boltzmann machines. pages 807–814, 2010.

[6] Diganta Misra. Mish: A self regularized non-monotonic activation function. *arXiv preprint arXiv:1908.08681*, 2019. URL `https://arxiv.org/abs/1908.08681`.

[7] OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. URL `https://arxiv.org/abs/2303.08774`.

[8] Qwen Team. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL `https://arxiv.org/abs/2505.09388`.