scientific reports

Check for updates

OPEN

Cross-cancer survival prediction using machine learning models

Lucas Buk Cardoso[1]✉, Jones Eduardo Egydio[1], Tatiana Natasha Toporcov[2], Nanci Yumi Utida[2], Maria Paula Curado[3], Gisele Aparecida Fernandes[3], Adeylson Guimarães Ribeiro[4], Bryan Gilvaz Chin[1] & Vanderlei Cunha Parro[1]

Among the many challenges faced by healthcare systems, cancer remains one of the most urgent. This makes the application of artificial intelligence a critical tool for enhancing early detection and optimizing treatment strategies, especially given the growing volume of patient data being collected. In this study, machine learning models trained on data for a specific type of cancer were employed to predict three-year survival after diagnosis for other cancer types. Two groups were considered: the most frequent cancers and those related to the digestive system. The data were extracted from the Hospital Based Cancer Registries of São Paulo, covering 2000 to 2019, with a consistent selection protocol across all cancer types to enable cross-prediction. XGBoost and LightGBM algorithms were used, choosing the best-performing model for predictions across different topographies. Using a combined dataset of oral cavity, esophageal, and stomach cancers, the model achieved a balanced accuracy of 80.18%, compared with 79.92% for the stomach-specific model. Statistical testing showed no significant difference between these values, suggesting comparable predictive performance. These results illustrate the potential of cross-prediction, especially for rare cancer types where data scarcity represents a significant challenge.

Keywords Survival prediction, Cancer, Machine learning, Cross-prediction, XGBoost

With rapid technological advancements, artificial intelligence (AI) has emerged as a transformative force capable of revolutionizing global healthcare[1]. Among the myriad challenges in healthcare, cancer remains one of the most significant, highlighting the critical need for accurate disease prognosis prediction. Machine learning (ML) tools offer great potential to address the multifaceted challenges of cancer survival prediction[2–4]. By leveraging advanced ML models and real-world clinical parameters, notable predictive accuracy has been achieved[2,5]. These models integrate diverse input parameters, capturing complex relationships across cancer types[6,7].

Registries, databases, and repositories have played pivotal roles in advancing AI applications in cancer treatment[8]. While modern AI techniques have demonstrated substantial efficacy across medical domains, their dependence on large datasets remains a key limitation[9,10]. Therefore, existing databases, such as national cancer registries, provide invaluable clinical data and serve as essential resources for research[8,11].

However, access to extensive registry databases is not universally available, especially for rare cancer subtypes and considering regional disparities in the quality of patient registries. To address this, transfer learning represents a great strategy – especially used in deep learning – where previously trained algorithms are used on other problems, reusing the training weights[12,13]. This technique is widely used in image data, but there is a lack of studies that verify its validity in tabular datasets, mainly due to the difficulty in finding high-quality and extensive databases[14], most of them are small and specialized, making it difficult to create complex models[15].

Even when trained on heterogeneous datasets, ML models can achieve good results in predictions[16], highlighting the importance of investigating approaches such as cross-prediction in areas with data limitations. The Hospital Based Cancer Registries of the state of São Paulo (RHC/SP), which has data on cancer patients since 2000[17], represents a great opportunity to validate cross-prediction across different types of cancer.

This study aimed to develop machine learning algorithms that use data from specific cancer types to predict three-year survival following diagnosis in different cancer topographies presents in RHC/SP. Following initial training, the models were independently evaluated for distinct cancer types to assess cross-prediction potential. This approach focused on two groups: the most frequent cancers and those related to the digestive system.

[1]Center for Embeded Eletronic Systems, Instituto Mauá de Tecnologia, São Caetano do Sul 09580-900, Brazil. [2]Epidemiology Department, Faculdade de Saúde Pública da Universidade de São Paulo, São Paulo 01246-904, Brazil. [3]Epidemiology and Statistics on Cancer Group, A.C. Camargo Cancer Center, São Paulo 01525-001, Brazil. [4]Information and Epidemiology, Fundação Oncocentro de São Paulo, São Paulo 05409-012, Brazil. ✉email: lucas.cardoso@maua.br

Results

Population characteristics

Among the most frequent cancer types, Colorectal cancer exhibited comparable proportions of patients across stages II, III, and IV (the same numbers for the digestive system, as the same dataset was used). Lung cancer demonstrated a predominance of stage IV cases, representing over 50% of patients. Breast and Prostate were predominantly diagnosed at stage II, with nearly 60% of Prostate cancer cases concentrated in this stage. Cervical cancer showed a balanced distribution across stages I, II, and III, with approximately 30% of patients in each category. In contrast, over 80% of Skin cancer cases were identified at stage I (Supplementary Fig. S1).

For digestive system topographies, stage IV predominated in Oral Cavity (53.7%), Oropharyngeal (72.8%), and Stomach cancers (47.5%). Esophageal and Small Intestine cancers exhibited a high prevalence of advanced stages, with approximately 30% of cases each in stages III and IV. Anus cancer diverged from this pattern, showing the highest proportions of patients in stages II and III (Supplementary Fig. S2).

Kaplan-Meier curves for the most frequent types show that Lung cancer has the worst prognosis 5 years after diagnosis, while Prostate, Breast, and Skin cancer have the highest survival rates (Supplementary Fig. S3). Regarding the types of the digestive system, the worst prognosis is for Esophagus cancer patients. Colorectal and Anus cancers have the highest 5-year survival rates after diagnosis (Supplementary Fig. S4).

Most frequent

Phase 1

It is important to note that, among the specialist models for the most frequent tumors, the LightGBM algorithm demonstrated superior performance in four of the six topographies analyzed – Breast, Skin, Lung, and Cervical (Supplementary Table 1). In contrast, for Prostate and Colorectal cancers, XGBoost performed better.

Feature analysis revealed that the clinical staging variable was critical in five of the six most frequent types for the most important columns for each specific cancer type (Supplementary Table 2).

After preparing the data for the most frequent types, the individual models were trained and validated, and predictions were made for the remaining types. Those with accuracy rates above 65% for both classes (0 – non-survival and 1 – survival) were selected, as shown in Supplementary Fig. S5. The highlighted cross-predictions were between Lung and Cervical, and between Cervical with Breast and Lung. These combinations were tested in Phase 2.

Phase 2

The results using the combination of types that showed strong predictive performance in Phase 1 for the most frequent cancer topographies are presented in Table 1. We selected balanced accuracy and f1-score to capture overall model performance with reduced sensitivity to class imbalance, while giving greater importance to accuracy in the cross-prediction analyses. The confusion matrices for the Phase 2 and Phase 3 tests for the most frequent types are shown in Supplementary Figs. S6 to S8.

By combining the types that showed cross-prediction potential identified in Phase 1, none of the results exceeded the accuracy of the specialist models for the respective types.

Phase 3

To conclude the tests for the most frequent types, the six topographies were combined into a single database to train a unified model (Table 1). However, the results obtained were inferior compared to especialist models when predicting each type individually, as observed in previous tests, demonstrating that there are no advantages with cross-prediction in this case.

Digestive system

Phase 1

When analyzing the specialist models for digestive system tumors, LightGBM algorithm demonstrated superior performance in four of the seven possible topographies: Oral Cavity, Oropharynx, Esophagus, and Stomach. In contrast, for Colorectal (which shares the same characteristics as the tests with the most frequent types), Small Intestine, and Anus cancers, XGBoost algorithm achieved better performance (Supplementary Table 1). The most important features for each specific digestive system cancer type, derived from the best-performing model,

	Specialist		Lung/Cervical		Cervical/Breast/Lung		All	
Type	BACC	f1-score	BACC	f1-score	BACC	f1-score	BACC	f1-score
Breast	0.7957	0.7345	-	-	0.7707	0.7704	0.7896	0.7639
Skin	0.7075	0.6576	-	-	-	-	0.7012	0.6779
Prostate	0.7455	0.6588	-	-	-	-	0.7246	0.7110
Colorectal	0.7641	0.7626	-	-	-	-	0.7333	0.7182
Lung	0.7865	0.6812	0.7443	0.7443	0.6301	0.6700	0.6006	0.6342
Cervical	0.7444	0.7416	0.6859	0.6830	0.7407	0.7274	0.7230	0.6972

Table 1. Balanced accuracy (BACC) and f1-score for the most frequent types, displaying the results of the specialist models, as well as the tests with the following combinations: lung and cervical; cervical, breast and lung; and all types combined.

are presented in Supplementary Table 3. Notably, the clinical staging variable emerged as the most important feature in six of the seven topographies.

After preparing the data for the digestive system types, individual machine learning models were trained, validated, and used to predict outcomes for other types. The predictions with accuracy rates above 65% for both classes, obtained from cross-predictions among digestive system topographies, were selected, as demonstrated in Supplementary Fig. S9. The highlighted cross-predictions include: Oral Cavity with Stomach and Colorectal; Esophagus with Oral Cavity and Stomach; Stomach with Colorectal; and Small Intestine with Oral Cavity and Stomach. These combinations were tested in Phase 2.

Phase 2

As in Phase 2 for the most frequent types, the topographies that demonstrated relevant cross-predictions were integrated into a unified database, enabling the training of a machine learning model to individually predict these highlighted types. Table 2 presents the results of all tests, with 95% confidence intervals for the metrics that outperformed the specialist models. The confusion matrices for the Phase 2 and 3 tests for the digestive system types are displayed in Supplementary Figs. S10 to S14.

For the combination of Esophagus, Oral Cavity, and Stomach, the predictions were inferior compared to the specialist models for the first two topographies. However, for Stomach, the results were numerically superior, reaching accuracy above 80%, in contrast to the lower values observed in the individual model. The f1-score showed a similar trend, rising from 0.7685 for the specialist model to 0.7718 for the combined model. McNemar's test yielded a p-value of 0.166, this demonstrates that the difference in accuracy between the models is not statistically significant; therefore, both performed similarly.

For the types Small Intestine, Oral Cavity, and Stomach, the balanced accuracy for Small Intestine was higher than that of the specialist model, although the prediction for the classes was unbalanced (Supplementary Fig. S12). Finally, in the tests with Stomach and Colorectal cancers, the predictions for colon and rectal cancer achieved higher accuracy and f1-score than the specialist model, with unbalanced classes. The McNemar test yielded a p-value < 0.05, indicating the difference in accuracy between the models was not statistically significant; therefore, no improvement in predictive performance can be claimed.

The other combinations with cross-prediction potential identified in Phase 1 do not exceeded the accuracy of the specialist models for the respective types.

Phase 3

Finally, all seven types were consolidated into a single dataset and used to train a model (Table 2). When predicting each topography individually, the Oral Cavity and Small Intestine types showed a superior balanced accuracy compared to specialists models – in the case of the f1-score there was only an increase for the Oral Cavity –, although the predictions for the two classes were unbalanced (Supplementary Fig. S14). This highlights the potential for cross-prediction in these cases.

The McNemar test for balanced accuracy improvements yielded p-values of 0.003 for Oral Cavity and 0.775 for Small Intestine cancer, demonstrating that the difference in accuracy between the models is statistically significant only in the former case.

Discussion

Based on the results obtained for three-year survival prediction after diagnosis, a marginal improvement over the specialist model was observed for Stomach cancer. When combining data from patients with Oral Cavity, Esophagus, and Stomach cancers into a single dataset, the model achieved a balanced accuracy of 80.18%, sensitivity of 80.06%, and specificity of 80.30%. In contrast, the model trained exclusively on Stomach cancer data yielded lower performance, with all metrics below 80%. McNemar's test did not indicate statistical significance in the difference between the predictions, suggesting the models have similar performance.

Superior numerical results compared to the specialist models were also achieved for Oral Cavity, Small Intestine, and Colorectal cancers. However, it is important to note that in these cases, performance improved for one class (survival or non-survival) but declined for the other, when compared to their respective specialized models, with p-value higher than 0.05 only for Small Intestine – not indicating statistical significance in the difference between the predictions –, so the models showed similar performance. Thus, it was possible to obtain good predictions for certain types of cancer using models trained with data from other topographies, with results being close to those obtained by specialized models. This highlights the potential of the cross-prediction approach when using the machine learning algorithms tested in this study.

This capability validated through cross-prediction – applied for the first time to RHC/SP data in this study – is particularly crucial for rarer cancer subtypes, in which the scarcity of well-annotated cases makes it impractical to train specialist models with sufficient robustness. By enabling models trained on more prevalent cancers to generate clinically meaningful predictions in data-limited settings, cross-prediction can help bridge a critical diagnostic and prognostic gap. This enhanced generalizability not only offers actionable insights that can support clinical decision-making when expert models are unavailable, but also broadens the applicability of computational oncology tools. Consequently, it represents a promising strategy to enhance disease characterization, inform therapeutic decision-making, and improve prognostic precision across a broader spectrum of malignancies, including tumor types historically underserved by data-driven methods. This technique of reusing model weights has been successful in predicting image-based datasets[18–20], but some studies show its successful application in case-limited studies[21,22].

Moreover, although various studies have applied machine learning to predict cancer survival, many are constrained by small sample sizes, where results may be affected by statistical bias and instability. For instance, a study on Oral Cavity cancer included 3,841 patients[23], while research on Breast cancer used samples of

Type	Specialist		Oral Cavity/ Stomach/ Colorectal		Esophagus/Oral Cavity/Stomach		Small Intestine/ Oral Cavity/ Stomach		Stomach/Colorectal		All	
	BACC	f1-score	BACC	f1-score	BACC (CI 95%)	f1-score (CI 95%)	BACC	f1-score	BACC (CI 95%)	f1-score (CI 95%)	BACC (CI 95%)	f1-score (CI 95%)
Oral Cavity	0.7440	0.7393	0.7362	0.7382	0.7240	0.7030	0.7417	0.7303	-	-	0.7495 (0.7387, 0.7605)	0.7473 (0.7366, 0.7582)
Oropharynx	0.6640	0.6255	-	-	-	-	-	-	-	-	0.6205	0.6322
Esophagus	0.7164	0.6091	-	-	0.6560	0.6722	-	-	-	-	0.5915	0.6184
Stomach	0.7992	0.7685	0.7624	0.7724	0.8018 (0.7916, 0.8124)	0.7718 (0.7619, 0.7821)	0.7936	0.7794	0.7532	0.7679	0.7744	0.7741
Small Intestine	0.7018	0.7001	-	-	-	-	0.7019	0.6854	-	-	0.7021 (0.6565, 0.7467)	0.6942 (0.6489, 0.7393)
Colorectal	0.7641	0.7626	0.7632	0.7660	-	-	-	-	0.7669 (0.7596, 0.7735)	0.7692 (0.7619, 0.7758)	0.7548	0.7579
Anus	0.7276	0.7266	-	-	-	-	-	-	-	-	0.7096	0.7092

Table 2. Balanced accuracy (BACC) and f1-score for the types related to digestive system, displaying the results of the specialist models, as well as the tests with the following combinations: oral cavity, stomach and colorectal; esophagus, oral cavity and stomach; small intestine, oral cavity and stomach; stomach and colorectal; and all types combined. 95% confidence intervals are shown for metrics that outperformed the specialist models.

only 210 patients[24]. Other studies reported even smaller cohorts, such as 44 patients for Lung cancer[25], 427 for Oropharyngeal cancer[26], and 75 for Stomach cancer[27]. In contrast, a Prostate cancer study featured a substantially larger cohort of 42,470 patients[28]. This considerable variability in sample size demonstrates the need for innovative strategies, such as cross-prediction, which can leverage knowledge acquired from high-data-volume datasets to improve survival predictions in data-scarce scenarios.

A major strength of this study lies in the use of the Hospital Based Cancer Registries of São Paulo, which represents a valuable resource for training machine learning algorithms. This database includes over one million registries since 2000, covering a diverse patient population in the state of São Paulo, and contains relevant clinical and epidemiological information such as age, education level, place of residence, treatment and tumor characteristics. As an example of the quality and completeness of the database, we can cite 97.5% of cases with microscopic confirmation and only 4.5% of patients without disease staging[17]. All these characteristics offer an excellent opportunity to test cross-prediction between different types of cancer.

Finally, among the limitations of this study, it is noteworthy that cancer types exhibit morphological heterogeneity, which can negatively affect the performance of models and cross-prediction tests when representing each type with a single dataset or mixing different topographies in a single database. Alternatively, cross-prediction could be used only for the less aggressive subtypes or tested on a single type and its respective characteristics. Additionally, the models were not validated on data from other regions; their generalizability was assessed only across different topographies within the RHC/SP. Another relevant point was the adoption of overall survival as an outcome – also considering deaths unrelated to cancer – which may introduce a competing risks bias, but was justified by the purpose of estimating the general prognosis of patients.

Although cross-prediction across cancer types appears promising, it is not yet ready for clinical application. Future work should therefore prioritize external validation using independent cancer registries and multicenter cohorts to assess generalizability and transportability. Studies should also examine performance across hospitals with different capacities and different regions to identify operational limits and potential equity issues. Integrating molecular and genetic data with registry variables can improve predictive accuracy and biological interpretability, and the use of survival models and deep learning techniques could bring gains in estimates from the cross-prediction approach.

Methods

Study population

The data were obtained from the RHC/SP, managed by Fundação Oncocentro de São Paulo (FOSP)[17]. The registry spans cases recorded from January 2000 to December 2023, totaling 1,178,688 patients. This dataset includes sociodemographic information, tumor characteristics, treatments, and hospital specifications. These data were collected from 79 public and private hospitals, with 94.2% of these institutions classified as high-complexity oncology centers within the Sistema Único de Saúde (SUS).

Two groups of cancer topographies were selected from the RHC/SP database. The first group comprises the most frequent types: Skin, Breast, Prostate, Colorectal, Lung, and Cervical. The second group includes digestive system associated topographies: Oral Cavity, Oropharynx, Esophagus, Stomach, Small Intestine, Colorectal, and Anus. To exclude pandemic disruptions, only patients diagnosed between 2000 and 2019 were included, and all were residents of São Paulo, Brazil. The outcome analyzed was three-year survival, chosen as an intermediate follow-up period after diagnosis to minimize severe class imbalance across cancer types with shorter or longer typical survival. Table 3 summarizes the study cohort, including the total number of patients, outcome class distributions, and training/test set sizes.

		3-Year-Survival		Data	
Type (CID-O)	N	No	Yes	Train	Test
Breast (C50)	104,532	20,721	83,811	78,399	26,133
Skin (C44)	100,883	23,196	77,687	75,622	25,261
Prostate (C61)	83,455	13,476	69,969	62,583	20,862
Colorectal (C18-C20)*	56,111	25,092	31,019	42,083	14,028
Lung (C34)	37,005	32,025	4,980	27,753	9,252
Cervical (C53)	21,470	9,239	12,231	16,102	5,368
Oral Cavity (C00-C06)	23,935	14,136	9,799	17,951	5,984
Oropharynx (C10)	4,922	3,646	1,276	3,691	1,231
Esophagus (C15)	13,961	12,086	1,875	10,470	3,491
Stomach (C16)	27,770	20,207	7,563	20,827	6,943
Small Intestine (C17)	1,594	883	711	1,195	399
Anus (C21)	2,276	1,043	1,233	1,707	569

Table 3. Overview of dataset composition by cancer type. For each tumour entity, the table reports: (i) the total number of patients available (N), (ii) the number of patients in each class for 3-year-survival, and (iii) the number of samples allocated to the training and test sets. * Same information for digestive system.

Selection of variables

The study's output variable was derived using the most recent patient information and the time interval (in days) between diagnosis and the last recorded follow-up. The outcome was defined as follows: 0 for patients who did not survive three years post-diagnosis, and 1 for those who survived at least three years. Twenty-three variables were selected as predictors for the machine learning models. Table 4 summarizes these variables and their descriptions.

The data selection protocol was applied uniformly to both the most frequent cancer types and digestive system-related topographies, ensuring consistency in input columns across all groups and enabling cross-prediction tests. It is important to emphasize that this standard data selection strategy may reduce the differentiation of histological subtypes; therefore, the results were interpreted considering this limitation, and studies exploring stratified analyses by subtype are recommended. Key exclusion criteria included patients aged under 20 years, non-residents of São Paulo state, cases with clinical staging of 0, X, or Y, diagnoses after 2019, and patients without microscopic confirmation of the diagnosis, as depicted in Fig. 1. The complete data selection steps are described in the Supplementary Material.

Following data selection and feature removal, patients with less than three years of follow-up data were excluded, retaining only those with sufficient follow-up duration ($\geq$ 3 years). This ensured eligibility for outcome assessment and excluded individuals censored prior to the three-year threshold. The final cohort sizes for each outcome class, along with training and test set distributions, are presented in Table 3.

Construction of the models

The cancer types included in the study were selected through two criteria: (1) the most prevalent types in the RHC/SP registry – Skin, Breast, Prostate, Lung, Colorectal, and Cervical – and (2) those associated with the digestive system – Oral Cavity, Oropharynx, Esophagus, Stomach, Small Intestine, Colorectal, and Anus.

After training specialist models, using data exclusively from a single cancer type, their predictive performance was evaluated across all other studied cancer types, as outlined in Phase 1 of Fig. 2. This process identified topographies with cross-prediction potential, defined as those for which specialist models achieved strong performance ($\geq$ 65% accuracy in both classes) when predicting another cancer type within the group.

For topographies demonstrating strong cross-prediction capabilities, a dedicated evaluation was performed using a dataset comprising only these types, as depicted in Phase 2 of Fig. 2. For instance, if specialist models for types A and B exhibited accuracy $\geq$ 65% in both classes, their datasets were merged to train machine learning models with the combined data. This threshold was set at 65% because the worst specialist model, for oropharyngeal cancer, has a balanced accuracy of 66%. The resulting models were then tested separately on type A and type B test sets to evaluate performance improvements relative to the original specialist models.

A third analysis focused on predictions from models trained on aggregated datasets encompassing all types within each group, as illustrated in Phase 3 of Fig. 2. Unified datasets were created: one for the six most frequent

Column	Description
IDADE	Patient's age
SEXO	Patient's gender
IBGE	Code of patient's city of residence according to IBGE
CATEATEND	Diagnostic care category
DIAGPREV	Previous diagnosis and treatment
EC	Clinical staging
CIRURGIA	Treatment received in hospital = surgery
RADIO	Treatment received in hospital = radiotherapy
QUIMIO	Treatment received in hospital = chemotherapy
HORMONIO	Treatment received in hospital = hormone therapy
TMO	Treatment received in hospital = bone marrow transplant
IMUNO	Treatment received in hospital = immunotherapy
OUTROS	Treatment received in hospital = others
CONSDIAG	Difference in days between the consultation and diagnosis dates
DIAGTRAT	Difference in days between treatment and diagnosis dates
ANODIAG	Year of diagnosis
DRS	Regional health departments
RRAS	Regional health care networks
IBGEATEN	IBGE code of the healthcare institution where the patient was treated
HABILIT2	Hospital qualifications - Categories
ESCOLARI	Code for patient's education level
IBGE_idem_IBGEATEN	IBGE code equal to IBGEATEN code
presenca_rec	Presence of recurrence

Table 4. Detailed description of all input features. For each feature column, the table provides: (i) the feature name as used in the datasets; (ii) a concise definition of what the feature measures.

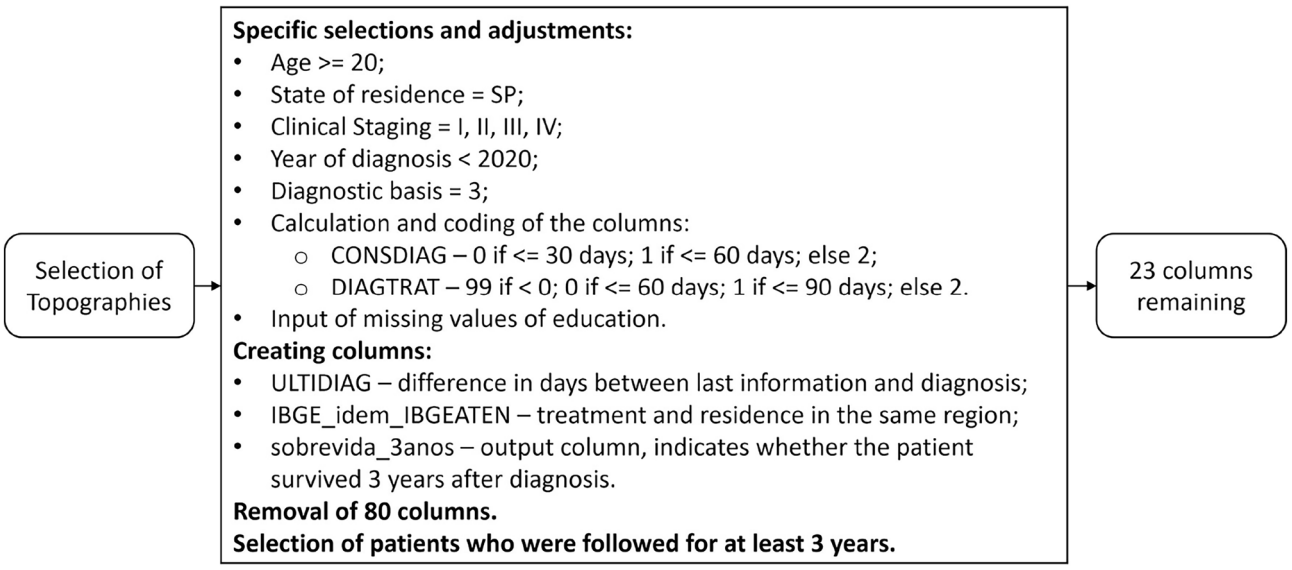

Fig. 1. The selection processes implemented to refine the databases for the chosen topographies ensured methodological consistency across all cancer types under study.

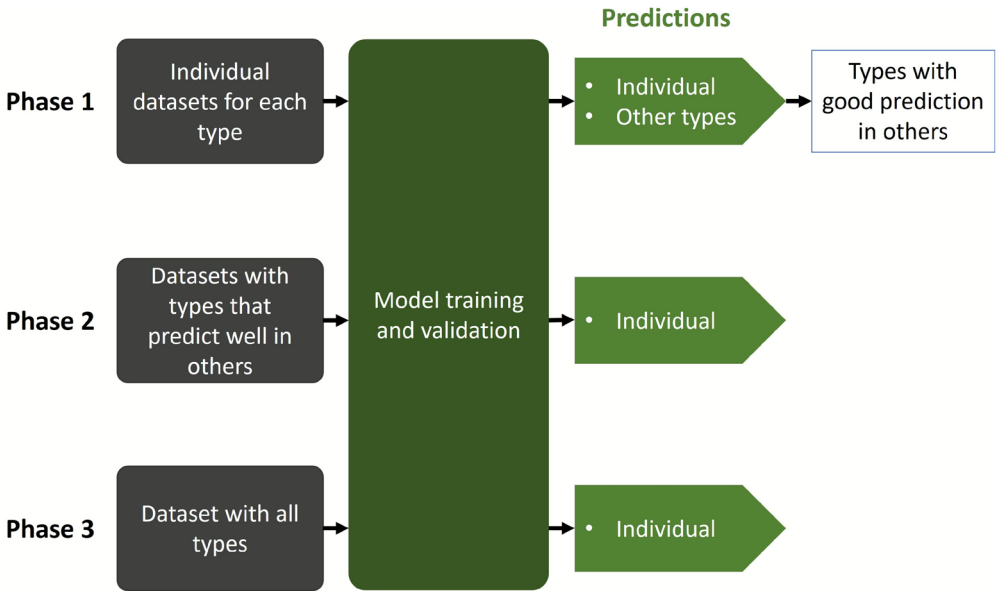

Fig. 2. The study comprised three phases. Phase 1 involved dedicated databases and specialist models for each cancer type. Topographies whose specialist models achieved predictive accuracy $\geq$ 65% in both outcome classes for other cancer types were selected for Phase 2. In Phase 2, types demonstrating strong inter-topography predictive performance were combined into a unified dataset to train a general model. This general model was then individually tested for each participating topography. In Phase 3, all cancer types were aggregated into a single dataset, and individualized predictions were generated for each topography within both the most frequent cancers and digestive system groups.

cancers and another for the seven digestive system topographies. Models were trained on these combined datasets, and individualized predictions were generated for each topography to compare against Phase 1 specialist models.

Two machine learning algorithms – XGBoost and LightGBM – were evaluated on the selected cancer topographies, as Gradient Boosting-based models performed better in survival predictions with RHC/SP compared to tree and deep learning models, as shown in a previous study by the research group[2]. Unlike specific survival models, the classification algorithms used in the study do not take censored data into account, which may influence survival estimates. Predictions were validated primarily through confusion matrices, while ROC curves and feature importance analyses were additionally derived for specialist models.

The XGBoost algorithm[29] employs gradient-boosted decision trees, sequentially combining weak learners to iteratively minimize residuals from prior trees. This approach enhances model robustness by focusing on misclassified instances, ultimately improving classification accuracy[30].

In contrast, LightGBM[31] optimizes gradient boosting through two key techniques: Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB). GOSS prioritizes data instances with larger gradients during training, while EFB reduces dimensionality by bundling mutually exclusive features (i.e., those rarely non-zero simultaneously)[30].

Hyperparameter optimization was performed using Optuna[32]. After selection for XGBoost and LightGBM, the model with the best performance was chosen to perform cross-validation tests.

Training and validation

The datasets for all types of cancer were split into two subsets – 75% for training and 25% for algorithm validation. Notably, the validation set used for cross-predictions remained consistent across all three phases and for all cancer types studied.

After the split, the training data underwent preprocessing, and the test data were transformed using ordinal encoding for categorical columns to convert category labels into numeric values[33], followed by z-score normalization to standardize all input features to a mean of 0 and a standard deviation of 1[33]. To correct imbalance in the output classes for specialist models, weights inversely proportional to the frequency of each class were assigned, so that the class with fewer examples received a higher weight during training and making the learning similar in both classes. Because class imbalance varies across cancer types, we avoided resampling and instead applied class weights to all expert models: weighting preserves the full dataset and facilitates cross-prediction, whereas oversampling – duplicating or synthetically generating minority examples – increases training size, computational cost and the risk of overfitting.

A summary of the metrics obtained by the best-performing model, selected from the two tested, is presented in Supplementary Table 1, comparing sensitivity, specificity, accuracy, and AUC (Area Under the Curve) values for training and testing. Likewise, the most important features using the SHAP method[34] for the specialist models are shown in Supplementary Tables 2 and 3.

To assess whether the improvement in metrics was statistically significant, it was used the McNemar test, which is appropriate for comparing two classifiers on the same dataset. This test evaluates whether the two types of disagreement – cases where model A is correct and model B is incorrect versus cases where model A is incorrect and model B is correct – occur at significantly different rates[35]. Thus, if the p-value < 0.05, the difference in accuracy between the models is statistically significant; otherwise, the models have similar performance.

Ethical considerations

Following the Lei Geral de Proteção de Dados Pessoais (LGPD) of Law No. 13,709, August 14, 2018, Section II – Processing of Sensitive Personal Data, as it is a research with a public database, which does not contain patients' personal data, approval from the Research Ethics Committee was not required.

Data availability

The raw database, with all types of cancer, are available in the FOSP website: https://fosp.saude.sp.gov.br/fosp/diretoria-adjunta-de-informacao-e-epidemiologia/rhc-registro-hospitalar-de-cancer/banco-de-dados-do-rhc/. The datasets generated and analysed, and the notebooks developed during the current study are available in the GitHub repository: https://github.com/CDIA-NSEE/cancer_cross_prediction.

Received: 10 June 2025; Accepted: 24 December 2025
Published online: 12 March 2026

References

1. Gaur, K. & Jagtap, M. M. Role of artificial intelligence and machine learning in prediction, diagnosis, and prognosis of cancer. *Cureus***14** (2022).
2. Buk Cardoso, L. et al. Machine learning for predicting survival of colorectal cancer patients. *Scientific reports***13**, 8874 (2023).
3. Montazeri, M., Montazeri, M., Montazeri, M. & Beigzadeh, A. Machine learning models in breast cancer survival prediction. *Technology and Health Care***24**, 31–42 (2016).
4. Kim, D. W. et al. Deep learning-based survival prediction of oral cancer patients. *Scientific reports***9**, 6994 (2019).
5. Zhao, R., Zhuge, Y., Camphausen, K. & Krauze, A. V. Machine learning based survival prediction in glioma using large-scale registry data. *Health informatics journal***28**, 14604582221135428 (2022).
6. Kalafi, E. et al. Machine learning and deep learning approaches in breast cancer survival prediction using clinical data. *Folia biologica***65**, 212–220 (2019).
7. Jung, J.-O. et al. Machine learning for optimized individual survival prediction in resectable upper gastrointestinal cancer. *Journal of Cancer Research and Clinical Oncology***149**, 1691–1702 (2023).
8. Wang, J. & Williams, M. Registries, databases and repositories for developing artificial intelligence in cancer care. *Clinical Oncology***34**, e97–e103 (2022).
9. Lee, C., Vogt, K. A. & Kumar, S. Prospects for ai clinical summarization to reduce the burden of patient chart review. *Frontiers in Digital Health***6**, 1475092 (2024).
10. Bohr, A. & Memarzadeh, K. *Artificial intelligence in healthcare* (Academic Press, 2020).
11. Esteva, A. et al. Dermatologist-level classification of skin cancer with deep neural networks. *nature***542**, 115–118 (2017).
12. Zhu, F. et al. Development and validation of a deep transfer learning-based multivariable survival model to predict overall survival in lung cancer. *Translational Lung Cancer Research***12**, 471 (2023).
13. Ayana, G., Dese, K. & Choe, S.-W. Transfer learning in breast cancer diagnoses via ultrasound imaging. *Cancers***13**, 738 (2021).
14. Gardner, J., Perdomo, J. C. & Schmidt, L. Large scale transfer learning for tabular data via language modeling. *Advances in Neural Information Processing Systems***37**, 45155–45205 (2024).
15. Hollmann, N. et al. Accurate predictions on small data with a tabular foundation model. *Nature***637**, 319–326 (2025).

16. Pettorruso, M. et al. Predicting outcome with intranasal esketamine treatment: A machine-learning, three-month study in treatment-resistant depression (esk-learning). *Psychiatry Research***327**, 115378 (2023).
17. Diretoria, F. O. S. P. *adjunta de informação e epidemiologia - banco de dados do rh*https://fosp.saude.sp.gov.br/fosp/diretoria-adjunta-de-informacao-e-epidemiologia/rhc-registro-hospitalar-de-cancer/banco-de-dados-do-rhc/
18. Jain, S. et al. Deep learning-based transfer learning for classification of skin cancer. *Sensors***21**, 8142 (2021).
19. Singh, R. et al. Imbalanced breast cancer classification using transfer learning. *IEEE/ACM transactions on computational biology and bioinformatics***18**, 83–93 (2020).
20. Wang, W., Li, Y., Yan, X., Xiao, M. & Gao, M. Breast cancer image classification method based on deep transfer learning In 190–197 (2024).
21. Chen, Q. et al. Transferability and interpretability of the sepsis prediction models in the intensive care unit. *BMC Medical Informatics and Decision Making***22**, 343 (2022).
22. Shao, D. et al. Rarenet: a deep learning model for rare cancer diagnosis. *Scientific Reports***15**, 22732 (2025).
23. Tan, J. Y. et al. Predicting overall survival using machine learning algorithms in oral cavity squamous cell carcinoma. *Anticancer Research***42**, 5859–5866 (2022).
24. Jang, W. et al. Artificial intelligence for predicting five-year survival in stage iv metastatic breast cancer patients: A focus on sarcopenia and other host factors. *Frontiers in Physiology***13**, 977189 (2022).
25. Miller, H. A., van Berkel, V. H. & Frieboes, H. B. Lung cancer survival prediction and biomarker identification with an ensemble machine learning analysis of tumor core biopsy metabolomic data. *Metabolomics***18**, 57 (2022).
26. Pan, X. et al. A survival prediction model via interpretable machine learning for patients with oropharyngeal cancer following radiotherapy. *Journal of Cancer Research and Clinical Oncology***149**, 6813–6825 (2023).
27. Akcay, M., Etiz, D. & Celik, O. Prediction of survival and recurrence patterns by machine learning in gastric cancer cases undergoing radiation therapy and chemotherapy. *Advances in radiation oncology***5**, 1179–1187 (2020).
28. Momenzadeh, N., Hafezalseheh, H., Nayebpour, M., Fathian, M. & Noorossana, R. A hybrid machine learning approach for predicting survival of patients with prostate cancer: A seer-based population study. *Informatics in Medicine Unlocked***27**, 100763 (2021).
29. Chen, T. & Guestrin, C. Xgboost. A scalable tree boosting system In 785–794 (2016).
30. Friedman, J. H. Greedy function approximation: a gradient boosting machine. *Annals of statistics* 1189–1232 (2001).
31. Ke, G. et al. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems* **30** (2017).
32. Akiba, T., Sano, S., Yanase, T., Ohta, T. & Koyama, M. Optuna. A next-generation hyperparameter optimization framework In 2623–2631 (2019).
33. Géron, A. Hands-on machine learning with Scikit-Learn, Keras. In *In and TensorFlow: Concepts, tools, and techniques to build intelligent systems ("* (O'Reilly Media, Inc", 2022).
34. Lundberg, S. M. et al. From local explanations to global understanding with explainable ai for trees. *Nature machine intelligence***2**, 56–67 (2020).
35. Dietterich, T. G. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation***10**, 1895–1923 (1998).

Author contributions
L.B.C., V.C.P., J.E.E., and T.N.T. conceived the study, L.B.C., J.E.E., and B.G.C. conducted the study, T.N.T., N.Y.U., M.P.C., G.A.F., and A.R. analysed the results. All authors reviewed the manuscript.

Funding
Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), Brazil. Process number: 2021/11794-4.

Declarations

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-34133-w.

Correspondence and requests for materials should be addressed to L.B.C.