scientific reports

Check for updates

OPEN

Survival prediction from imbalanced colorectal cancer dataset using hybrid sampling methods and tree-based classifiers

Sadegh Soleimani, Mahsa Bahrami & Mansour Vali✉

Colorectal cancer is a high mortality cancer, with a mortality rate of 64.5% for all stages combined. Clinical data analysis plays a crucial role in predicting the survival of colorectal cancer patients, enabling clinicians to make informed treatment decisions. However, utilizing clinical data can be challenging, especially when dealing with imbalanced outcomes, an aspect often overlooked in this context. This paper focuses on developing algorithms to predict 1-, 3-, and 5-year survival of colorectal cancer patients using clinical datasets, with particular emphasis on the highly imbalanced 1-year survival prediction task. We utilized a colorectal cancer dataset from the Surveillance, Epidemiology, and End Results (SEER) database, which exhibits high imbalance in the 1-year (1:10) survival analysis and an imbalance in the 3-year (2:10) analysis, achieving balance in the 5-year analysis. The pre-processing step consists of removing records with missing values and merging categories with less than 2% share for each categorical feature to limit the number of classes of each component. Edited Nearest Neighbor, Repeated Edited Nearest Neighbor (RENN), Synthetic Minority Over-sampling Technique (SMOTE), and pipelines of SMOTE and RENN approaches were used for balancing the data with tree-based classifiers, including Decision Tree, Random Forest, Extra Tree, eXtreme Gradient Boosting, and Light Gradient Boosting Machine (LGBM). The performance evaluation utilizes a 5-fold cross-validation approach. In the case of 1-year, our proposed method with LGBM significantly outperforms other sampling methods with the sensitivity of 72.30%. For the task of 3-year survival, the combination of RENN and LGBM achieves a sensitivity of 80.81%, indicating that our proposed method works best for highly imbalanced datasets. Additionally, when predicting 5-year survival, the sensitivity reaches 63.03% using LGBM. Our proposed method significantly improves mortality prediction for the minority class of colorectal cancer patients. RENN followed by SMOTE yields better sensitivity in the classifiers, with LGBM as the predictor performing best for 1- and 3-year survival. In the 5-year task, LGBM outperforms other models in terms of F1-score.

Keywords Colorectal cancer, Repeated edited nearest neighbor, Synthetic minority over-sampling technique, Survival prediction, SEER

Colorectal cancer (CRC) is a prevalent form of cancer, ranking fourth in diagnoses and third in cancer-related deaths worldwide as of 2018[1,2]. Risk factors for developing CRC include smoking, obesity, unhealthy lifestyle, and alcohol consumption[3]. A confident prediction of each case's survivability would be helpful to health maintenance and can help physicians have a fair approximation of patients' survivability[4].This information helps in making better decisions for patients' treatment.

Clinical data analysis provides a unique opportunity to study the importance of collected features and analyze patients' data. Machine learning algorithms can be used to achieve this goal[5]. Despite the high prevalence and death rates of CRC worldwide, it seems to be overlooked in such analysis and research compared to breast cancer or lung cancer[6]. Clinical data processing involves several stages: collecting and cleaning patient data, secure and organized storage, analysis using statistical and machine learning methods, results interpretation, and dissemination to relevant parties[7]. The pre-processing stage involves imputing incomplete data, removing outliers, and normalizing data for further analysis. Feature selection aims to improve the performance of machine learning algorithms by selecting most discriminative features for the task, thus decreasing computational

Department of Biomedical Engineering, Faculty of Electrical Engineering, K. N. Toosi University of Technology, 16315-1355, 1631714191 Tehran, Iran. ✉email: mansour.vali@eetd.kntu.ac.ir

complexity and making algorithms more cost-efficient. Sampling and balancing techniques are mainly used to improve the capability of models to have better results on classification. Finally, machine learning algorithms are applied for decision making. Machine learning algorithms rely on balanced training data for accurate classification. However, CRC datasets are often highly imbalanced, with a significantly lower number of deceased patients compared to survivors. This imbalance can lead to poor sensitivity, misleading accuracy scores by biasing classifiers towards the majority class, and reducing generalizations. To mitigate these issues, resampling techniques such as oversampling and undersampling methods have been proposed. However, applying these methods indiscriminately may introduce noise or result in loss of valuable information.

Several automatic survival predictions from clinical data have been proposed using machine learning and statistical approaches. Wang et al.[8], proposed a 5-year survival model based on a Bayesian network for cancer patients with second primary cancer. By combining eleven cancer databases, they gathered 7,845 patients' data. They used synthetic minority over-sampling technique (SMOTE) to balance the dataset. Then artificial neural networks, support vector machines (SVM), logistic regression (LR), and the proposed method used to report results. The results showed that SMOTE boosts the sensitivity, and the proposed method outperforms other classifiers.

Among different machine learning algorithms, trees have shown that they can be robust in many situations[9]. And since clinical data are related to biological systems, they are complex[10], they should be treated with robust classifiers to ensure decent results, or the results should be compared with tree-based classifiers[11]. Tree-based classifiers are also more robust against imbalance in data[12]. It was observed that ensemble tree-based classifiers show better results in dealing with imbalance classification[13].

Gao et al.[14], employed various data mining methods to predict colorectal cancer outcomes using the National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) dataset. In their study, all of the methods were compared with TNM staging (the progression of cancer is classified based on a combination of the parameters Tumor (T), Node (N), and Metastasis (M)). Feature selection of each classifier was made using either backward step-wise feature selection or a genetic algorithm. Their study showed that the decision tree (DT) algorithm was the most accurate in predicting the survival rate of colorectal cancer patients. Similarly, Another study[15], proposed a two-stage model based on tree ensembles to predict the survival of patients with advanced-stage colorectal cancer. At this stage survival data were asymmetric due to the survival probability of 10%; therefore, a model was designed to provide satisfactory results despite this asymmetry. The classification model was trained using multiple symmetric models, with samples randomly chosen from the larger class to match the smaller class size. The final classification was determined by voting. Agrawal et al.[16] applied an ensemble data mining approach for survival prediction in lung cancer patients from the SEER dataset. The ensemble voting classifier in this paper achieved the best performance for 5-year survival prediction. The study showed that the ensemble approach improved the accuracy of prediction compared to individual models. A similar approach was taken by Al-Bahrani et al.[17], who used ensemble data mining along with SMOTE for balancing to develop a colon cancer survival prediction model.

Recent advancements in deep learning for survival analysis highlight a variety of approaches to enhance interpretability and address complex data types. In another paper by Al-Bahrani et al.[18], by using a deep neural network and by finding an optimal number of layers from 1 to 8, it obtained acceptable results compared to the two widely used random forest (RF) and LR methods. This paper used the data of 188,000 colon cancer patients from the SEER dataset and the best Area Under the Curve (AUC) was achieved with five hidden layers at about 0.88, showing its potential as a reliable method for predicting survivability in colon cancer patients. The development of SurvSHAP(t) extends Shapley Additive exPlanations (SHAP) to accommodate survival data, offering feature importance scores that account for both event occurrence and time[19]. Similarly, SurvLIME adapts the Local Interpretable Model-agnostic Explanations (LIME) framework for survival data, providing localized explanations of individual predictions by considering censoring and time-to-event information[19]. These methods enhance model transparency, making it easier for researchers to interpret the contributions of various features to survival outcomes. In contrast, recent methods like Dynamic DeepHit and DeepCompete have introduced novel architectures to tackle the challenges of time-varying covariates and competing risks. Dynamic DeepHit extends the original DeepHit model by incorporating time-varying features, thus improving the model's ability to predict survival outcomes in longitudinal studies[20]. On the other hand, DeepCompete utilizes neural ordinary differential equations to model competing risks, providing a sophisticated framework for handling multiple risk types simultaneously[21]. These advancements represent significant strides in adapting deep learning techniques to complex survival analysis tasks, balancing predictive accuracy with interpretability.

In Valentini et al.[22] article, they used a dataset from five European clinical trials for rectal cancer to develop models and nomograms to predict local recurrence, distant metastasis, and 5-year survival. Models and nomograms were based on Cox regression. The c-index for local recurrence, distant metastasis, and 5-year survival was 0.68, 0.73, and 0.70, respectively. Bowles et al.[23] used a Cox regression method for calculating rectal cancer conditional survival. A dataset consisting of 22,610 patients with rectal adenocarcinoma from the SEER database was used in this paper. It was shown that 5-year survival time in all stages except stage I significantly improved. In Momenzadeh et al.[24], authors proposed a hybrid machine learning approach for predicting the survival of patients with prostate cancer data from SEER database. They applied a cluster centroid under-sampling approach to balance the data. The results showed that eXtreme Gradient Boosting (XGBoost) outperformed other classifiers in binary classification, while SVM achieved the best performance in the three-class type. Wang et al.[25], applied a multivariate Cox proportional hazards survival model for rectal cancer patients. In a recent study by Wu et al.[26], the integration of Cox proportional hazards regression (CoxPH) with advanced machine learning algorithms for survival analysis was studied. The study explored random survival forest, gradient boosting, extra survival trees, and survival trees, highlighting their superior performance in handling high-dimensional data and right-censored survival data compared to traditional methods. These advanced models combine survival information with complex algorithms, making them powerful tools for time-to-event predictions. While CoxPH is widely used and reliable, it may not process high-dimensional data as efficiently, underscoring the advantage of these advanced machine learning models in providing more accurate and comprehensive survival analysis.

With the development of the field of machine learning in survival analysis and predictive modeling, researchers have reviewed the use of optimization techniques combined with the goal of improving classification accuracy and effectiveness of feature selection. Current research has demonstrated the advantage of incorporating these optimization algorithms with machine learning methods, hence leading to better disease classification and predictive modeling in healthcare. For example, the coupling of Greylag goose optimization with a multilayer perceptron has been shown to optimize lung cancer classification through effective feature selection and improvement in diagnostic precision[27]. Likewise, the optimized Binary Particle Swarm Optimization model was used for forecasting COVID-19 spread with better predictive precision due to effective feature selection and adaptive modeling of epidemiological data[28]. Furthermore, a framework for feature selection utilizing a snake optimization algorithm was established to facilitate the swift identification of cardiovascular disease, thereby minimizing computational complexity and improving early diagnostic potential[29].

Further advancements extend these innovative approaches to other critical areas of disease prediction. One study introduced the hybrid Optimized Gradient Boosting (framework for hepatitis C virus disease prediction, integrating hybrid optimization techniques with gradient boosting to tailor intervention strategies and improve predictive performance[30]. Another approach employed an enhanced Convolutional Neural Network-Long Short-Term Memory deep learning model for the early detection of a disease, demonstrating the potential of combining spatial and temporal feature learning for robust disease forecasting[31]. Together, these studies underscore the promising impact of novel optimization algorithms and deep learning architectures in addressing diverse diagnostic challenges across both human and plant health.

While these methods and researches offer significant advancements, many suffer from limitations such as sensitivity to class imbalance, reliance on extensive feature engineering, or computational inefficiencies. In this study, we address the challenge of imbalanced survival prediction in CRC by proposing a novel hybrid sampling technique that combines the strengths of oversampling and undersampling while mitigating their weaknesses. Our approach ensures a fair and comprehensive analysis of CRC survival prediction, with a focus on improving sensitivity and robustness across multiple time horizons. The proposed method is particularly beneficial for cases where severe class imbalance affects predictive performance, as observed in the 1-year and 3-year survival tasks. Our proposed approach has the following contributions:

- Process data using statistical feature selection techniques along with correlogram analysis.
- We investigate a novel sampling technique (RENN + SMOTE) for balancing the dataset without losing valuable patterns.
- Compare various tree-based classifiers within a unified framework for 1-, 3-, and 5-year CRC survival prediction.
- Compare the impact of different sampling techniques on the performance of classifiers, particularly in handling highly imbalanced and imbalanced datasets.

Experimental setup

A block diagram of our proposed method is shown in Fig. 1. Colorectal cancer data were pre-processed, then different balancing methods were used to overcome imbalanced data problems. Finally, different tree-based machine-learning methods were applied for classification.

Dataset and pre-processing

We used the SEER database[32] gathered since 1973. Colon and rectum cancer records from 2010 to 2015 were imported. The age of patients at the time of diagnosis was asserted, and those who were not in the range of 18 to 85 were removed from the dataset. With these enhancements total of 103,885 records, we removed those who died of anything but colorectal cancer. Moreover, the study is focused on the adenocarcinoma form of cancer, so from histologic type ICD-O-3, only the adenocarcinoma type was selected, and others were removed.

Furthermore, records with missing values, coded or labeled Unknown, were also dropped. As a result of these modifications, the dataset size shrank to 42,764 records, and unbalancing in the data increased from 1:5 to 1:10 due to a high rate of missing values in patients who died under one year. This also indicates that missing values could be informative if filled with a tag, and the model would observe that pattern. This article chose the dropping policy instead of imputing missing values. The remaining records consist of about 8% with Grade I, 74% with Grade II, 15% with Grade III, and only 2% with Grade IV. Histologic Type consists of 78% *Adenocarcinoma*, *Not Otherwise Specified (NOS)*, 10% Adenocarcinoma in tubulovillous adenoma and others were put into the Others category. Sigmoid Colon occupies the most share by 21.5%, and the second and third most common places for cancer tumors in the dataset are Cecum and Rectum, NOS, respectively.

Selected variables can be seen in Table 1. All variables were either categorical or were transformed into categories; for example, regional nodes examined is the total number of regional lymph nodes that were removed and examined; however, it was changed into five classes – 0, 1-9,10–20, 20–30, and > 30 Node Examined.

Furthermore, some features were maintained to reduce their bins or simplify, like Median Household Income, whose categories were merged to half of its primary number of classes.

We set our overall survival to 1-,3-, and 5-year, and since all data has enough follow-ups for this task, we map survival months to Survived and Not-Survived based on the task. A histogram of survival months of patients is shown in Fig. 2. As we can see, the data is highly imbalanced at 1-year survival, so about 90% of the records are labeled as Survived and 10% labeled as Not Survived.

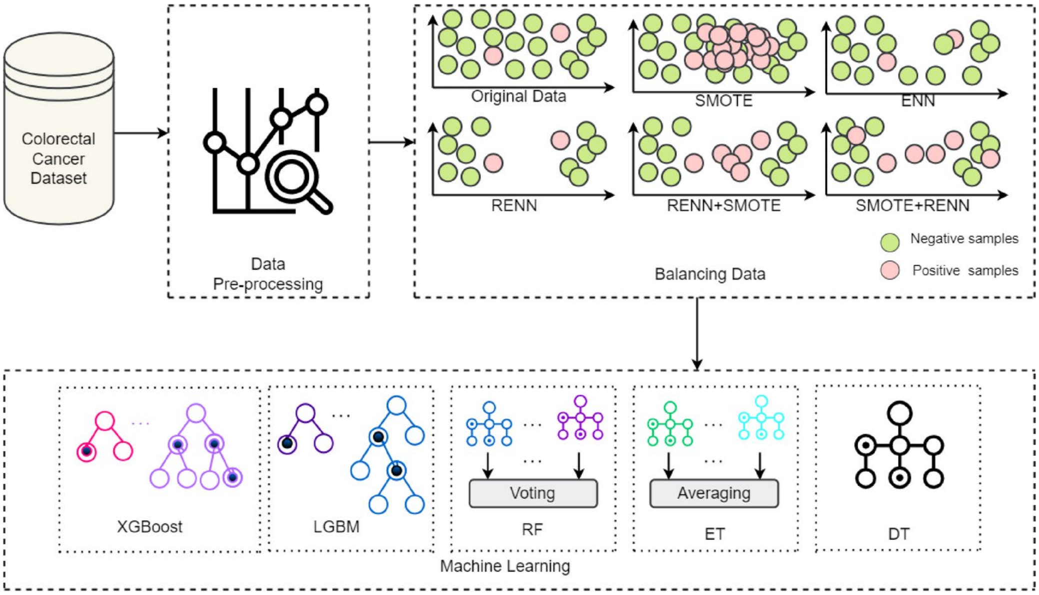

Fig. 1. Block diagram representation of the proposed framework for colorectal cancer survival prediction.

Feature Name	Number of Categories	Name of Categories
Sex	2	Male/Female
Grade	4	I/ II/ III/ IV
Race	4	W/B/AI/API
Primary Site	8	Sigmoid Colon/ Rectum/ Cecum/ Ascending Colon/ Rectosigmoid junction/ Transverse Colon/ Descending Colon/ Others
Histologic type ICDO-3	4	8140-8389
Surgery of Primary Site	3	No Surgery, Tumor Destruction, Tumor Resection
CS Tumor Size	5	No Tumor/ 1–20/ 20–40/ 40–60/ 60–80
Regional nodes examined	5	0/ 1–9/ 10–30/ 30/ Others
Regional nodes positive	3	0/ Any positive node/ Not Applicable
Median Household Income	5	< 40k/ 40k-50/50–60/60–70/ >70k
Perineural Invasion Recode	2	Not Present/ Present
AJCC T, 6th ed	5	Tis/ T1/T2/T3/T4
AJCC N, 6th ed	3	N0/N1/N2
AJCC M, 6th ed	2	M0/M1
CEA Pretreatment Interpretation Recode	3	Not Documented/ Positive/ Negative
CS extension	4	Categories are based on codes in the documentation of SEER
SEER Combined Mets at DX-liver	2	Positive/ Negative
CS lymph nodes	6	Categories are based on codes in the documentation of SEER
Number of in situ/malignant tumors	3	Categories are based on codes in the documentation of SEER

Table 1. Selected features from the colorectal cancer dataset.

After categorizing and reducing categories of all features, we performed an ANOVA test to remove non-significant elements from the dataset; in the result, Race, Sex, and Number of in situ/malignant tumors had a p-value higher than 0.05, while others had a p-value of less than 0.0001. As a result, non-significant features were removed.

We then performed a Crammer's V between the 16 remaining features to ensure that highly correlated features were not among the selected ones. The correlogram of the components is shown in Fig. 3.

In Fig. 3, most features correlate with each other except the component related to the income of the patient's family. This is due to the basis of the feature, which is socioeconomic, unlike other clinical features. Also, we can see that positive nodes and examined nodes have a high correlation because of their nature.

Correlations of features like the one between Collaborative Stage (CS), lymph nodes and AJCC 7 N should be considered as well. The N category is assigned a value of CS Lymph Nodes and the value of CS Site-Specific

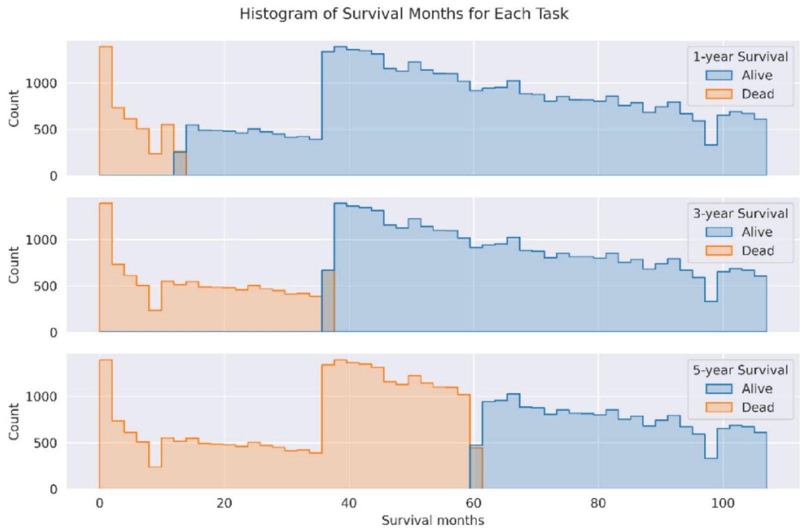

Fig. 2. Histogram of patient survival times (in months).

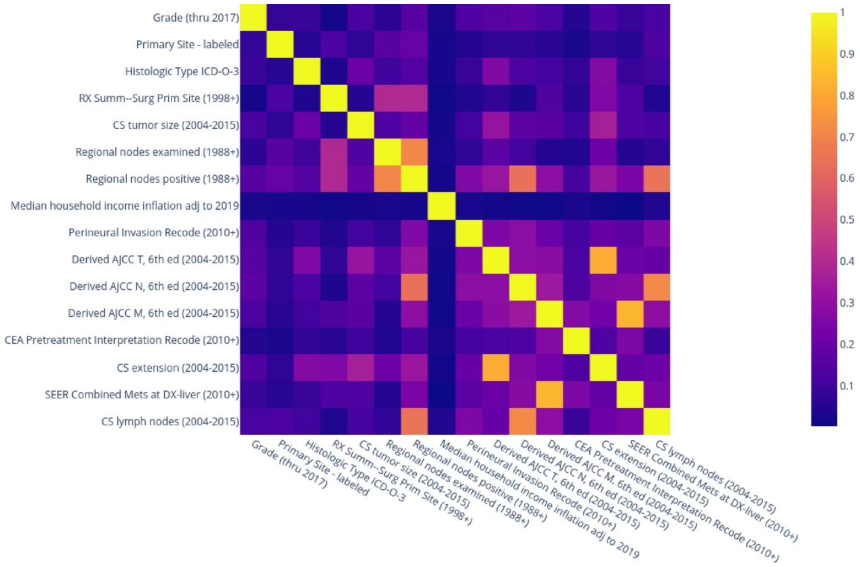

Fig. 3. Correlogram of features in the processed dataset.

Factor 3, the Number of Positive Ipsilateral Axillary Lymph Nodes, so we selected AJCC 7 N and removed the other component.

Balancing method

Imbalance data has a detrimental impact on machine learning algorithms. It biases classifiers into opting for the majority class as the classification's outcome, reducing precision or recall by misclassifying the minority class samples. Several methods are available to overcome this problem, including under-sampling and over-sampling. This study compared over-sampling and under-sampling methods, including Edited Nearest Neighbor (ENN)[33], Repeated Edited Nearest Neighbor (RENN)[34], and SMOTE[35], for data balancing. Also, we developed hybrid sampling approaches from SMOTE and RENN for balancing the data.

ENN under-samples the majority class by removing the samples with a different label than their k-nearest neighbors in the training set[33]. RENN repeats ENN algorithms until it draws all models with other brands in their k-nearest neighbors[34]. SMOTE oversamples the minority class by generating synthetic samples using a random sample of the minority class and its k-nearest neighbors[35]; however, since we have categorical features in the dataset, we have used a modified version that its output fits the categories of each class.

In highly imbalanced datasets, the challenge is simultaneously having good sensitivity and specificity, which can be reached by balancing the dataset or putting weights in the model to compensate for the gap in the class population. Here, we need better sensitivity in the models with no sampling method. We have tried to balance the dataset so that sensitivity is increased without a considerable loss in specificity.

In Figs. 4, 1000 samples were generated so that about 10% went to class A (purple) and 90% went to class B (yellow). And each of the samplers that we use in this paper were applied to those samples with 75% separability. Samplers' results are shown in Fig. 4. The number of samples varies based on the method. ENN drops 165 samples that are considered noisy by a five-neighbor metric. RENN removes more samples in comparison to ENN, which is plane. RENN reduces the dataset size to 728 from 1000, with about 100 more samples reduced than ENN. Unlike the under-sampling methods, SMOTE adds to the dataset size and increases the dataset's length to 1792 samples. The combined approach follows the RENN method by a SMOTE that reduces dataset size at first and then increases its size to 1248 samples. At last, SMOTE + RENN is shown which result in the same number of samples as the input.

A pseudocode representation of our proposed RENN + SMOTE sampling method is provided in Algorithm 1, detailing the step-by-step procedure for applying RENN followed by SMOTE to enhance class balance.

```
Procedure Hybrid_Sampling(X, y, k, α)
  // Step 1: Apply RENN for noise removal
  repeat
    Train k-NN classifier on (X, y)
    Predict labels ŷ for all instances in X
    Identify misclassified instances: M = {x_i | ŷ_i ≠ y_i}
    Remove all instances in M from (X, y)
  until no more instances are removed

  // Step 2: Apply SMOTE for minority class oversampling
  for each minority instance $x_i$ in X do
    Find k nearest neighbors from the minority class
    Randomly select a neighbor $x_j$
    Generate a synthetic instance $x_s$ using:
      $x_s = x_i + \lambda (x_j - x_i)$, where $\lambda \sim U(0,1)$
    Add $x_s$ to (X, y)
  until minority class reaches α * majority class size

  return (X', y')
End Procedure
```

Algorithm 1. Hybrid RENN + SMOTE Sampling

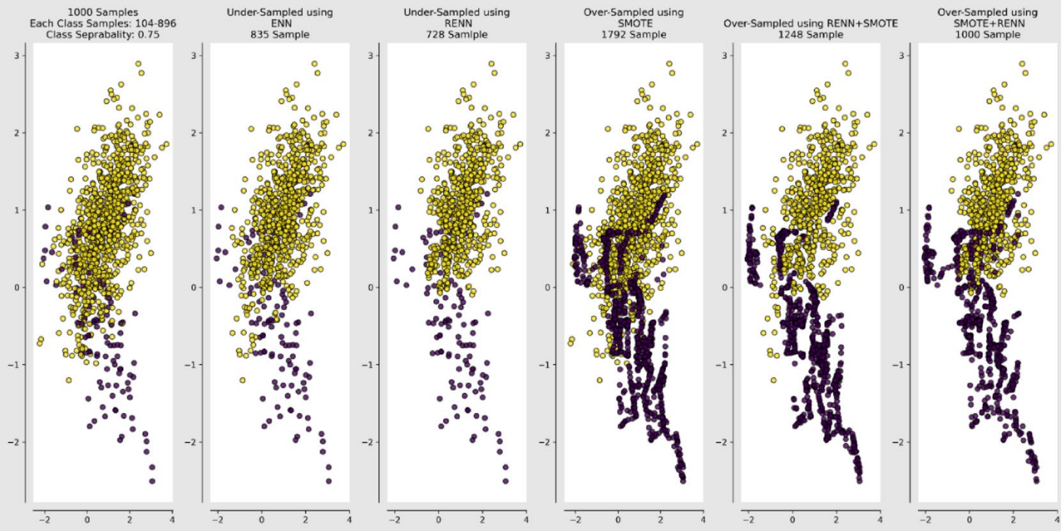

Fig. 4. Comparison of tailored sampling methods including ENN, RENN, SMOTE, RENN + SMOTE, and SMOTE + RENN techniques.

Looking at the details of samples in each sampler, we can see that ENN and RENN remove a significant amount of data from the overlap area of classes and biases the outcome of the classification. Also, SMOTE strengthens patterns that are weak by making and generating numerous samples from them and as a result models predict falsely based on these patterns. RENN + SMOTE removes noisy data and then creates synthesis data. SMOTE + RENN draws data patterns as RENN does and generates a bundle of data from minority class in the overlap.

Classification

Among different simple machine learning algorithms, the decision tree[36] is more interpretable for clinical data processing since it can recognize non-linear patterns more effectively than other base learners. A DT is a supervised learning technique that can be used for classification and regression problems, but it is mainly preferred for solving classification problems. It is a tree-based classifier where internal nodes represent the features of a dataset, branches represent the decision rules, and each leaf node represents the outcome. DTs usually mimic human thinking ability while making a decision, so it is easy to understand and can be beneficial for solving decision-related problems. However, it may have an overfitting issue, which can be resolved using ensemble methods.

Ensemble methods try to prevent overfitting and reduce the variance of prediction of its bias based on their structure. Extra trees (ET) and RF[37] are ensemble methods based on voting or averaging. RF creates numerous trees and trains each tree by a proportion on the train set. Then each tree estimates the input, and the RF prediction aggregates those predictions by averaging or voting. ET has the same structure, though it splits nodes based on random thresholds, resulting in even lower variance than RF—main extended DT versions, generally ensemble DT.

Boosting is an ensemble method where unused models are included to adjust the errors made by existing models. Models are included successively until no further advancements can be made. XGBoost[38] is a more advanced method that opts for a reduction in bias via reducing the error by building new trees and aggregating the results. Gradient boosting is an approach in which unused models foresee the residuals or blunders of earlier models and are added together to create the ultimate prediction. It is called gradient boosting since it employs a gradient descent algorithm to decrease the loss when including new models. LGBM[39] is considered a rapid algorithm and the foremost utilized algorithm in machine learning for getting fast and high-precision results. LGBM develops trees vertically, whereas other algorithms develop trees horizontally, meaning it grows tree leaf-wise, while different algorithms develop level-wise. It'll select the leaf with max delta loss to conceive. When raising the same leaf, a leaf-wise algrithm can diminish more upsets than a level-wise algorithm.

Results and discussion

For more reliable and accurate analysis, we developed a 5-fold cross-validation in which 80% of data was used for training, and 20% was used for testing the performance of algorithms in each fold. Several metrics, including accuracy, sensitivity, specificity, and F1-score, were computed as described in Eqs. (1–4) to evaluate the proposed method.

$$Accuracy = \frac{TN + TP}{TN + FP + TP + FN} \tag{1}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{2}$$

$$Specificity = \frac{TN}{TN + FP} \tag{3}$$

$$F1_{score} = \frac{1}{1 + \frac{FN}{2TP} + \frac{FP}{2TP}} \tag{4}$$

In Eqs. (1–4), TP, TN, FP, and FN are represented as a true positive, true negative, false positive, and false negative, respectively. The performance of the proposed method is summarized in Tables 2 and 3.

To evaluate whether the differences in classifier performance are statistically significant, we applied statistical hypothesis testing. A one-way ANOVA (Analysis of Variance) test was used to compare the mean performance of different sampling methods across classifiers. When significant differences were found, a post-hoc paired t-test with Bonferroni correction was conducted to determine pairwise differences between sampling strategies. Statistical significance was set at $p < 0.05$.

Table 2 reports the performance of models without any sampling method for all three tasks. 1-year survival prediction task is highly imbalanced; in this regard, the accuracy and specificity are about 90%, while the sensitivity and F1-score are near 10%, which means that the false negative rates are too high for an acceptable classification. The best sensitivity and F1-score for the 1-year survival prediction task were obtained using the XGBoost algorithm, which shows the robustness of boosting algorithm using trees to counter imbalance data compared to other classifiers used in this article.

The 3-year survival prediction data in Table 2 is still unbalanced, although milder than the 1-year survival prediction. The accuracy and specificity in this task are degraded by about 10% and 5%, respectively, while the sensitivity and F1-score are increased by about 30%. The best sensitivity and F1-score in this task were achieved by LGBM, which suggests that LGBM is a better classifier for mild imbalanced data classification among the used classifiers. The 5-year survival prediction was balanced, and the best accuracy, sensitivity, and F1-score were achieved by LGBM, which illustrates this classifier's strength in classifying clinical outcomes.

Model	Survival time	Accuracy	Sensitivity	Specificity	F1-score
DT	1-year	90.63%	6.74%	99.36%	11.96%
	3-year	81.91%	40.28%	94.14%	50.24%
	5-year	60.55%	58.10%	63.76%	62.57%
ET	1-year	90.61%	2.80%	99.57%	5.33%
	3-year	82.10%	38.06%	94.94%	48.98%
	5-year	60.94%	60.96%	60.92%	63.91%
RF	1-year	90.70%	7.30%	99.38%	12.88%
	3-year	82.16%	38.90%	94.77%	49.60%
	5-year	61.02%	62.09%	59.62%	64.38%
XGBoost	1-year	90.22%	10.67%	98.50%	17.06%
	3-year	81.57%	38.74%	94.05%	48.68%
	5-year	60.22%	62.49%	57.23%	64.06%
LGBM	1-year	90.62%	8.04%	99.21%	13.91%
	3-year	82.22%	40.66%	94.34%	50.80%
	5-year	61.16%	63.03%	58.70%	64.81%

Table 2. Performance comparison between different classifiers in the proposed 1-year, 3-year, and 5-year survival prediction.

Model	Sampling	Accuracy	Sensitivity	Specificity	F1-score
DT	SMOTE	71.39%	59.23%	72.66%	28.07%
	ENN	89.40%	24.59%	96.15%	30.44%
	RENN	82.72%	48.81%	86.25%	34.75%
	RENN + SMOTE	65.78%	70.99%	65.23%	28.11%
	SMOTE + RENN	80.97%	45.91%	84.63%	31.27%
ET	SMOTE	78.29%	54.00%	80.82%	31.91%
	ENN	89.88%	23.15%	96.82%	30.12%
	RENN	84.31%	47.00%	88.19%	36.09%
	RENN + SMOTE	68.87%	70.49%	68.70%	29.91%
	SMOTE + RENN	83.35%	42.61%	87.59%	32.54%
RF	SMOTE	78.33%	49.26%	81.36%	30.00%
	ENN	89.40%	26.58%	95.93%	32.09%
	RENN	82.66%	51.56%	85.90%	35.92%
	RENN + SMOTE	67.57%	70.89%	67.23%	29.18%
	SMOTE + RENN	82.97%	39.75%	87.47%	30.56%
XGBoost	SMOTE	76.13%	49.90%	78.86%	28.26%
	ENN	88.59%	28.09%	94.89%	31.70%
	RENN	81.48%	51.46%	84.60%	34.37%
	RENN + SMOTE	60.91%	71.25%	59.83%	25.57%
	SMOTE + RENN	79.68%	44.32%	83.37%	29.14%
LGBM	SMOTE	76.24%	50.59%	78.90%	28.64%
	ENN	88.91%	29.70%	95.07%	33.55%
	RENN	78.73%	57.59%	80.93%	33.79%
	RENN + SMOTE	63.06%	72.30%	62.10%	26.95%
	SMOTE + RENN	82.34%	41.27%	86.62%	30.58%

Table 3. Performance comparison between different sampling models and classifiers in the proposed 1-year survival prediction.

In Fig. 5, sensitivity and F1-score of all methods are demonstrated. As we can see in Fig. 5, the results of 1-year survival are poor and by having more balance in the data, they get better. Also, variations in 1-year task's results are higher between classifiers in compare to other tasks.

In Table 3, the results of the 1-year survival classification are reported in terms of accuracy, sensitivity, specificity, and F1-score. This paper compares SMOTE, ENN, RENN, and combinations of RENN and SMOTE as sampling methods using different tree-based algorithms.

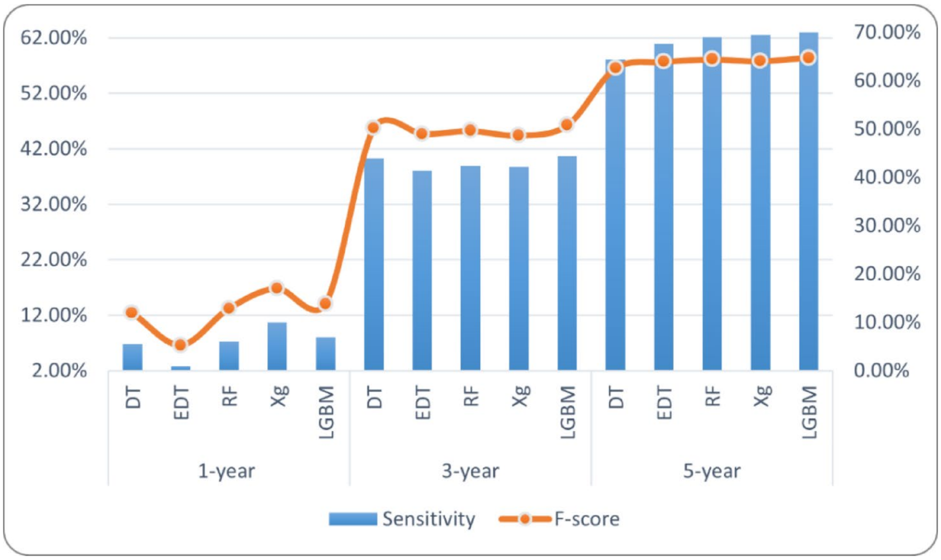

Fig. 5. Sensitivities (left axis) and F1-score (right axis) of all tasks for the classifiers with no sampling method.

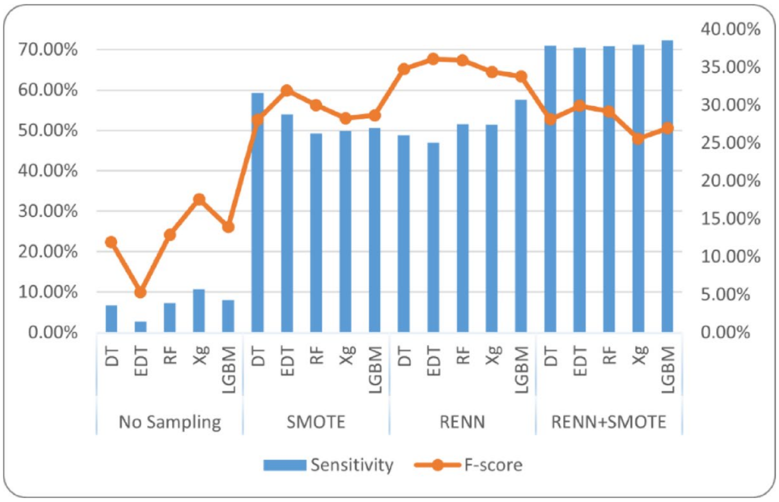

Fig. 6. Sensitivity and F1-score of 1-year survival for each classifier using No-sampling and 3 top samplers, including SMOTE, RENN, and RENN + SMOTE techniques.

In 1-year survival, the best sensitivity is achieved by a RENN and SMOTE orderly pipeline with LGBM, resulting in 15% higher sensitivity than RENN with LGBM. The sampling method results in the best sensitivity in other classifiers as well. This indicates that the process is robust for increasing the true positive rate in the model. Also, the models' sensitivity increased significantly compared to no-sampling models of Table 2, which are less than 11% for 1-year survival—in 1-year survival combining SMOTE and RENN with LGBM, which had a sensitivity of about 0.08 in the non-sampling method, improved its capability of predicting minority class 9 times better, reaching 0.72. This significant improvement is also nearly 15% higher sensitivity than SMOTE or RENN sampling method alone. This indicates the model's capability to deal with rare event predict imbalanced datasets. Also, the procedure is robust on all tree-based classifiers, showing a range of 0.12–0.27 increase in F1-score using different methods. Sensitivities and F1-score of three top sampling methods including SMOTE, RENN, and RENN + SMOTE comparing with no sampling method are shown in Fig. 6, and as we can see the significant improvement in the sampling methods is quite notable. Moreover, the significant rise in the sensitivity does not decrease F1-score. The highest sensitivity belongs to RENN + SMOTE across classifiers which has a slight lower F1-score in compare to RENN. This less improvement in F1-Score is a good tradeoff for about 0.2 rise in sensitivity.

Table 4 reports models with sampling methods for 3-year survival. In 3-year survival, the best sensitivity is achieved using RENN as the sampler and LGBM as the classifier. The increase in sensitivity is about 40%, double the true positive rate with no sampler, showing that minority class would be predicted more accurately

Model	Sampling	Accuracy	Sensitivity	Specificity	F1-score
DT	SMOTE	74.38%	65.05%	77.10%	53.41%
	ENN	79.86%	54.37%	87.29%	54.92%
	RENN	68.87%	74.71%	67.20%	52.00%
	RENN + SMOTE	68.89%	74.67%	67.21%	52.00%
	SMOTE + RENN	80.27%	48.14%	89.64%	52.41%
ET	SMOTE	76.83%	64.12%	80.53%	55.54%
	ENN	80.69%	55.10%	88.15%	56.29%
	RENN	68.14%	77.42%	65.43%	52.31%
	RENN + SMOTE	68.88%	76.77%	66.58%	52.69%
	SMOTE + RENN	80.60%	49.52%	89.66%	53.54%
RF	SMOTE	76.74%	63.08%	80.72%	55.04%
	ENN	79.88%	58.65%	86.07%	56.82%
	RENN	65.63%	79.53%	61.58%	51.09%
	RENN + SMOTE	66.00%	79.01%	62.21%	51.20%
	SMOTE + RENN	80.28%	49.53%	89.25%	53.14%
XGBoost	SMOTE	75.48%	61.42%	79.58%	53.07%
	ENN	78.90%	57.91%	85.01%	55.33%
	RENN	64.79%	78.21%	60.87%	50.06%
	RENN + SMOTE	64.99%	78.29%	61.11%	50.24%
	SMOTE + RENN	78.55%	49.36%	87.06%	50.95%
LGBM	SMOTE	76.62%	63.00%	80.59%	54.89%
	ENN	79.13%	61.27%	84.35%	57.00%
	RENN	63.44%	80.81%	58.38%	49.95%
	RENN + SMOTE	63.83%	80.65%	58.93%	50.16%
	SMOTE + RENN	79.78%	50.61%	88.29%	53.06%

Table 4. Performance comparison between different sampling models and classifiers in the proposed 3-year survival prediction.

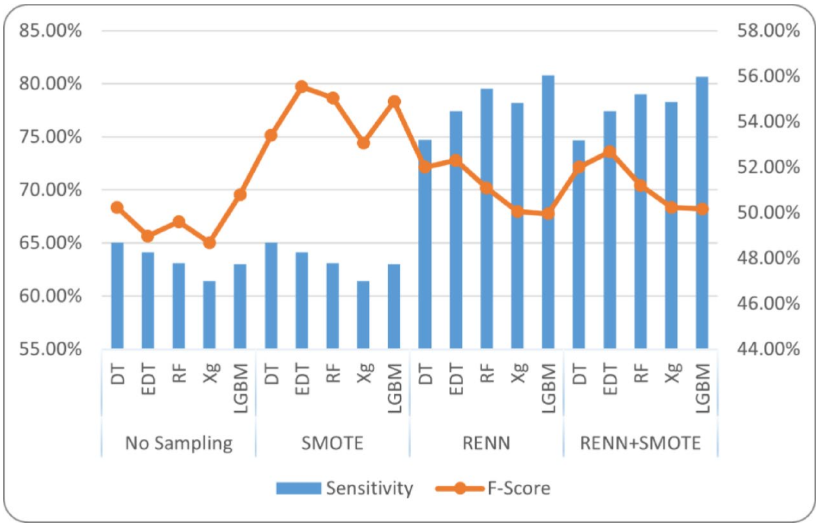

Fig. 7. Sensitivity and F1-score of 3-year survival for each classifier using No-sampling and 3 top samplers, including SMOTE, RENN, and RENN + SMOTE techniques.

by sampling. Also, the F1-score is not significantly dropped; even we can see an increase in the results of the models that RF using ENN as a sampler has a considerably higher F1-score than models with no sampling. As we can see in Fig. 7, the performance of RENN and the hybrid sampler does not differ, this is due to the fact that 3-year survival is much less imbalance in compare to 1-year task. The structure of the proposed method focuses on highly imbalanced datasets, as a result the metrics for it in 3-year survival are not better than RENN.

The ANOVA comparison of various sampling methods indicates that there exist statistically significant differences among the methods across all the four performance metrics: Accuracy (F = 101.98, $p < 0.00001$), Sensitivity (F = 276.20, $p < 0.00001$), Specificity (F = 126.46, $p < 0.00001$), and F1-score (F = 67.22, $p < 0.00001$). These observations imply that the selection of a sampling method does have a strong impact on classifier performance.

The Post-hoc Tukey's HSD test subsequently determines the precise locations of these differences. The RENN + SMOTE method consistently demonstrates significant gains in Sensitivity, outperforming all other sampling methods. In particular, in comparison with No Sampling, RENN + SMOTE boosts Sensitivity by 64.07% points ($p < 0.00001$), indicating its strong capacity for recall enhancement. Furthermore, it outperforms both ENN ($p < 0.00001$) and RENN ($p < 0.00001$) significantly, validating the assertion that this method is extremely effective at increasing the detection of positive instances.

At F1-score level, RENN + SMOTE also shows significant enhancements over No Sampling (+ 15.72 points, $p < 0.00001$) and presents significant differences regarding RENN (-7.04, $p = 0.0004$), which means that it enhances the precision-recall trade-off. However, regarding ENN and SMOTE, differences are not significant, meaning that while RENN + SMOTE achieves good results, its F1-score enhancements are less pronounced.

A noteworthy trade-off is found in Specificity space, with RENN + SMOTE having a considerable decline in performance relative to No Sampling (-34.59, $p < 0.00001$) and ENN (-31.15, $p < 0.00001$). This reveals that although RENN + SMOTE boosts sensitivity, it does so at the cost of producing more false positives, thereby lowering the performance of the model in accurately classifying negative instances.

Finally, the use of ANOVA with post-hoc tests enhances the validity of the results, validating that RENN + SMOTE significantly helps to enhance Sensitivity and F1-score at the expense of Specificity. The foregoing results emphasize the need for the selection of a resampling approach to be aligned with the overall classification objectives—when recall takes precedence, RENN + SMOTE is an effective approach in handling imbalanced datasets.

As the results show, there is a significant improvement in sensitivity and F1-score when using the RENN + SMOTE sampling strategy; however, this comes at a computational cost. The model trained with this sampling method exhibits a notably higher training time. While no sampling method allows for training all classifiers in under one second, the proposed method results in approximately four minutes of training. Despite this increase in training time, the model remains valuable, especially given the strong computational power available today.

Conclusions

The field of survival prediction for colorectal cancer has seen significant advancements, with researchers proposing methods and critical features for assessing the survivability of this disease. Clinical data along with machine learning plays a pivotal role in analyzing colorectal cancer, providing invaluable insights into patients' conditions. However, the imbalance of disease outcomes poses a challenge for machine learning models, as they often struggle to handle such imbalances. In our study, we built models for predicting 1-, 3-, and 5-year survival of colorectal cancer from SEER data, and compared the sampling models and tree-based classifiers with each other.

We investigated the use of oversampling techniques for robust predictions; however, minority oversampling alone may not sufficiently improve prediction outcomes due to the high volume of synthetic data created. Conversely, under-sampling may result in the loss of valuable training data while attempting to balance the classes. To overcome these limitations, we employed the RENN algorithm to remove noisy data, followed by the SMOTE to generate synthetic data and balance the classes. The proposed novel approach yielded a notable enhancement in sensitivity while demonstrating minimal impact on other metrics such as F1-score, thus establishing its suitability as a sampling technique.

One of the key findings of this study is the trade-off between sensitivity and specificity when using different sampling methods. Oversampling techniques like SMOTE improve sensitivity by increasing the representation of the minority class; however, this often comes at the cost of specificity, as the classifier may misclassify some majority class instances. Conversely, undersampling techniques such as ENN and RENN enhance specificity by removing noisy data but can reduce sensitivity by discarding valuable minority class instances. Our hybrid approach (RENN + SMOTE) effectively balances this trade-off, as evidenced by the results in the 1-year survival task, where it achieved one of the highest sensitivity scores across models while maintaining reasonable specificity. However, in some cases, the increased computational cost of hybrid sampling might not be justified for mildly imbalanced datasets, as observed in our 3-year survival prediction scenario.

Our study underscores the significance of addressing data imbalance in survival prediction for colorectal cancer. Through the integration of SMOTE and RENN, we achieved improved sensitivity in predicting 1-year survival. However, it should be noted that due to its higher computational cost, this method is not recommended for mildly imbalanced datasets. This indicates that using the proposed method in highly imbalanced datasets can be invaluable, while in cases of mild imbalance, simpler techniques may be preferable.

Future studies can explore additional techniques such as handling missing values, feature selection, dimension reduction, and utilizing alternative machine learning methods to further enhance predictive performance. Specifically, comparing different missing value handling methods, including imputing them with median, iterative imputer, and nearest neighbor imputer, would be valuable for future investigations. Additionally, further research can explore the effects of different hybrid sampling ratios to optimize the sensitivity-specificity trade-off based on dataset characteristics.

Data availability

The dataset is extracted from SEER dataset (www.seer.cancer.gov) which is available for research purposes on request in the SEER application. Also, you can email our corresponding author (mansour.vali@eetd.kntu.ac.ir) for the dataset we used specifically.

Received: 7 November 2024; Accepted: 14 April 2025
Published online: 25 April 2025

References

1. Parkin, D. M., Bray, F., Ferlay, J. & Pisani, P. *Global cancer Stat. 2002 CA: cancer J. Clin.*, **55**, 2, 74–108, (2005).
2. Rawla, P., Sunkara, T. & Barsouk, A. Epidemiology of colorectal cancer: incidence, mortality, survival, and risk factors. *Gastroenterol. Review/Przegląd Gastroenterologiczny*. **14** (2), 89–103 (2019).
3. Botteri, E. et al. Smoking and colorectal cancer: a meta-analysis, *Jama*, vol. 300, no. 23, pp. 2765–2778, (2008).
4. Bilimoria, K. Y., Stewart, A. K., Winchester, D. P. & Ko, C. Y. The National Cancer data base: a powerful initiative to improve cancer care in the united States. *Ann. Surg. Oncol.* **15**, 683–690 (2008).
5. Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V. & Fotiadis, D. I. Machine learning applications in cancer prognosis and prediction. *Comput. Struct. Biotechnol. J.* **13**, 8–17 (2015).
6. Keum, N. & Giovannucci, E. Global burden of colorectal cancer: emerging trends, risk factors and prevention strategies. *Nat. Reviews Gastroenterol. Hepatol.* **16** (12), 713–732 (2019).
7. Hjollund, N. H. I., Valderas, J. M., Kyte, D. & Calvert, M. J. Health data processes: a framework for analyzing and discussing efficient use and reuse of health data with a focus on patient-reported outcome measures. *J. Med. Internet. Res.* **21** (5), e12412 (2019).
8. Wang, K. M., Wang, K. J. & Makond, B. Survivability modelling using bayesian network for patients with first and secondary primary cancers. *Comput. Methods Programs Biomed.* **196**, 105686 (2020).
9. Gardner, J., Popović, Z. & Schmidt, L. Subgroup Robustness Grows On Trees: An Empirical Baseline Investigation, *arXiv preprint arXiv:2211.12703*, (2022).
10. Zitnik, M. et al. Machine learning for integrating data in biology and medicine: principles, practice, and opportunities. *Inform. Fusion*. **50**, 71–91 (2019).
11. Afrash, M. R., Mirbagheri, E., Mashoufi, M. & Kazemi-Arpanahi, H. Optimizing prognostic factors of five-year survival in gastric cancer patients using feature selection techniques with machine learning algorithms: a comparative study. *BMC Med. Inf. Decis. Mak.* **23** (1), 54 (2023).
12. Batista, G. E., Prati, R. C. & Monard, M. C. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsl.* **6** (1), 20–29 (2004).
13. Ghiasi, M. M., Zendehboudi, S. & Mohsenipour, A. A. Decision tree-based diagnosis of coronary artery disease: CART model. *Comput. Methods Programs Biomed.* **192**, 105400 (2020).
14. Gao, P. et al. Which is a more accurate predictor in colorectal survival analysis? Nine data mining algorithms vs. the TNM staging system. *PLoS One*. **7** (7), e42015 (2012).
15. Wang, Y. et al. A tree ensemble-based two-stage model for advanced-stage colorectal cancer survival prediction. *Inf. Sci.* **474**, 106–124 (2019).
16. Agrawal, A., Misra, S., Narayanan, R., Polepeddi, L. & Choudhary, A. Lung cancer survival prediction using ensemble data mining on SEER data. *Sci. Program.* **20** (1), 29–42 (2012).
17. Al-Bahrani, R., Agrawal, A. & Choudhary, A. Colon Cancer Survival Prediction Using Ensemble Data Mining on SEER Data, IEEE, pp. 9–16 (2013).
18. Al-Bahrani, R., Agrawal, A. & Choudhary, A. Survivability prediction of colon cancer patients using neural networks. *Health Inf. J.* **25** (3), 878–891 (2019).
19. Krzyziński, M., Spytek, M., Baniecki, H. & Biecek, P. SurvSHAP (t): time-dependent explanations of machine learning survival models. *Knowl. Based Syst.* **262**, 110234 (2023).
20. Utkin, L. V., Satyukov, E. D. & Konstantinov, A. V. SurvNAM: the machine learning survival model explanation. *Neural Netw.* **147**, 81–102 (2022).
21. Cho, H. J., Shu, M., Bekiranov, S., Zang, C. & Zhang, A. Interpretable meta-learning of multi-omics data for survival analysis and pathway enrichment, *Bioinformatics*, vol. 39, no. 4, p. btad113, (2023).
22. Valentini, V. et al. Nomograms for predicting local recurrence, distant metastases, and overall survival for patients with locally advanced rectal cancer on the basis of European randomized clinical trials. *J. Clin. Oncol.* **29** (23), 3163–3172 (2011).
23. Bowles, T. L. et al. An individualized conditional survival calculator for patients with rectal cancer. *Dis. Colon Rectum*. **56** (5), 551 (2013).
24. Momenzadeh, N., Hafezalseheh, H., Nayebpour, M. R., Fathian, M. & Noorossana, R. A hybrid machine learning approach for predicting survival of patients with prostate cancer: A SEER-based population study. *Inf. Med. Unlocked*. **27**, 100763 (2021).
25. Wang, S. J. et al. An interactive tool for individualized Estimation of conditional survival in rectal cancer. *Ann. Surg. Oncol.* **18**, 1547–1552 (2011).
26. Wu, Y. et al. Survival prediction in second primary breast cancer patients with machine learning: an analysis of SEER database, (in eng). *Comput. Methods Programs Biomed.* **254**, 108310. https://doi.org/10.1016/j.cmpb.2024.108310 (Sep 2024).
27. Elkenawy, E. S. M., Alhussan, A. A., Khafaga, D. S., Tarek, Z. & Elshewey, A. M. Greylag Goose optimization and multilayer perceptron for enhancing lung cancer classification. *Sci. Rep.* **14** (1), 23784 (2024).
28. Alkhammash, E. H. et al. Application of machine learning to predict COVID-19 spread via an optimized BPSO model, *Biomimetics*, vol. 8, no. 6, p. 457, (2023).
29. Tarek, Z., Alhussan, A. A., Khafaga, D. S., El-Kenawy, E. S. M. & Elshewey, A. M. A snake optimization algorithm-based feature selection framework for rapid detection of cardiovascular disease in its early stages. *Biomed. Signal Process. Control.* **102**, 107417 (2025).
30. Elshewey, A. M. et al. Optimizing HCV disease prediction in Egypt: The hyOPTGB framework, *Diagnostics*, vol. 13, no. 22, p. 3439, (2023).
31. Alzakari, S.A., Alhussan, A.A., Qenawy, AS.T. et al. Early detection of potato disease using an enhanced convolutional neural network-long short-term memory deep learning model. *Potato Res.* **68**, 695–713 (2025).
32. Howlader, N. et al. Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer. gov) SEER* Stat Database: Incidence-SEER 18 Regs Research Data + Hurricane Katrina Impacted Louisiana Cases, Nov 2015 Sub (2000–2013) < Katrina/Rita Population Adjustment>-Linke. *Rita population adjustment¿ e Linke*, 2015. (2015).
33. Wilson, D. L. Asymptotic properties of nearest neighbor rules using edited data. *IEEE Trans. Syst. Man. Cybernetics*. **no. 3**, 408–421 (1972).
34. Tomek, I. An experiment with the edited nearest-neighbour rule. *IEEE Trans. Syst., Man, Cybernetics*. **6**. 448–452 (1976).
35. Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **16**, 321–357 (2002).

36. Breiman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J. Classification and regression trees (Wadsworth, Belmont, CA), *ISBN-13*, pp. 978-0412048418, (1984).
37. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
38. Chen, T. et al. Xgboost: extreme gradient boosting, *R package version 0.4-2*, vol. 1, no. 4, pp. 1–4, (2015).
39. Ke, G. et al. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems* **30**, 3149–3157 (2017).

Author contributions

Sadegh Soleimani: Responsible for data fetching, preprocessing, statistical analysis, drafting the article, cover letter, prepared figures, and collaborating on implementation strategies.Mahsa Bahrami: Implemented the machine learning models and balancing techniques, contributed ideas for the article structure, prepared figures, and assisted in arranging the article.Mansour Vali: Provided feedback and suggestions for improving the work, requested enhanced visuals and content, and contributed ideas for improving the manuscript's clarity and impact.

Declarations

Competing interests

The authors declare no competing interests.

Human studies/informed consent

The authors carried out no human studies for this article.

Animal studies

The authors carried out no animal studies for this article.

Additional information

Correspondence and requests for materials should be addressed to M.V.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.