BMC Public Health

RESEARCH

Open Access

Comparison of spatial prediction models from Machine Learning of cholangiocarcinoma incidence in Thailand

Oraya Sahat[1], Supot Kamsa-ard[2*], Apiradee Lim[3], Siriporn Kamsa-ard[2], Matias Garcia-Constantino[4] and Idongesit Ekerete[4]

Abstract

Background Cholangiocarcinoma (CCA) poses a significant public health challenge in Thailand, with notably high incidence rates. This study aimed to compare the performance of spatial prediction models using Machine Learning techniques to analyze the occurrence of CCA across Thailand.

Methods This retrospective cohort study analyzed CCA cases from four population-based cancer registries in Thailand, diagnosed between January 1, 2012, and December 31, 2021. The study employed Machine Learning models (Linear Regression, Random Forest, Neural Network, and Extreme Gradient Boosting (XGBoost)) to predict Age-Standardized Rates (ASR) of CCA based on spatial variables. Model performance was evaluated using Root Mean Square Error (RMSE) and R^2 with 70:30 train-test validation.

Results The study included 6,379 CCA cases, with a male predominance (4,075 cases; 63.9%) and a mean age of 66.2 years (standard deviation = 11.1 years). The northeastern region accounted for most of the cases (3,898 cases; 61.1%). The overall ASR of CCA was 8.9 per 100,000 person-years (95% CI: 8.7 to 9.2), with the northeastern region showing the highest incidence (ASR = 13.4 per 100,000 person-years; 95% CI: 12.9 to 13.8). In the overall dataset, the Random Forest model demonstrated better prediction performance in both the training (R^2 = 72.07%) and testing datasets (R^2 = 71.66%). Regional variations in model performance were observed, with Random Forest performing best in the northern, northeastern regions, while XGBoost excelled in the central and southern regions. The most important spatial predictors for CCA were elevation and distance from water sources.

Conclusion The Random Forest model demonstrated the highest efficiency in predicting CCA incidence rates in Thailand, though predictive performance varied across regions. Spatial factors effectively predicted ASR of CCA, providing valuable insights for national-level disease surveillance and targeted public health interventions. These findings support the development of region-specific approaches for CCA control using spatial epidemiology and machine learning techniques.

Keywords Cholangiocarcinoma, Spatial Predictions, Prediction Models, Machine Learning, Population-based cancer registries, Thailand

*Correspondence:
Supot Kamsa-ard
supot@kku.ac.th
Full list of author information is available at the end of the article

Background

Cholangiocarcinoma (CCA), a malignant tumor originating from the biliary epithelium, represents a public health challenge in Thailand [1]. While relatively rare worldwide, CCA exhibits an exceptionally high incidence in Southeast Asia, with the northeastern region of Thailand reporting the highest global rates (85 per 100,000 person-years) [2]. This striking geographic disparity is primarily attributed to the prevalence of the *Opisthorchis viverrine (O. viverrini)* infection, although other risk factors such as hepatolithiasis, primary sclerosing cholangitis, praziquantel treatment for *O. viverrini* [3–6], and hepatitis B and C [7–9] also contribute to the disease burden [8]. The unique distribution of CCA in Thailand underscores the need for sophisticated spatial analysis to better understand and address this critical health issue.

Spatial epidemiology plays a crucial role in elucidating disease patterns and their underlying causes [10]. In the context of CCA in Thailand, spatial prediction models offer valuable insights into the complex interplay of environmental (particularly proximity to water source) [11–13] and biological factors that influence disease distribution. These models can significantly enhance resource allocation for screening and treatment, enable targeted public health interventions, and deepen our understanding of risk factors [14].

Traditional statistical approaches to spatial epidemiology, while valuable, often struggle to capture the complex non-linear relationships and interactions between multiple environmental, demographic, and social factors that influence disease distribution. In recent years, Machine Learning has emerged as a powerful alternative for analyzing complex health data patterns [15]. Unlike conventional statistical methods, Machine Learning algorithms can identify intricate, non-linear relationships without requiring pre-specified model structures, making them particularly well-suited for spatial epidemiological research where relationships between variables may be complex and multifaceted. [16, 17].

The Machine Learning models demonstrate several advantages for on spatial epidemiology compared to traditional approaches. They offer superior predictive accuracy when analyzing complex, non-linear relationships in health data [18]. Algorithms such as Random Forests and Neural Networks can integrate diverse data sources, including satellite imagery, census data and environmental measurements, creating more comprehensive spatial predictions [19, 20]. These techniques also excel at handling large datasets with multiple variables and can identify patterns that might be missed by conventional statistical methods [21].

Our study aligns with Thailand's National Artificial Intelligence Strategy (NAIS) Action Plan for 2022–2027, with consists of five strategies aimed at national development through Artificial Intelligence (AI) applications. Strategy 4 specifically focuses on advancing intelligent technology systems using AI to create novel computational learning and reasoning approaches. The NAIS Action Plan leverages these intelligent systems across various sectors and supports research for the National Artificial Intelligence as a Service (AIaaS) Platform [22].

Previous studies on CCA in Thailand have been limited to short-term retrospective analyses covering only selected provinces or regions, without nationwide assessment. Additionally, most existing research has relied on traditional statistical prediction methods rather than advanced machine learning techniques. Therefore, this study aims to compare the performance of spatial prediction models using Machine Learning approaches to analyze CCA occurrence throughout Thailand. By conducting a comprehensive spatial analysis focus on demographic, environmental, and climatic variables, we can identify high-risk areas and potential causative factors. This mapping initiative can inform local public health strategies and provide valuable recommendations for CCA management and prevention, while contributing to the broader fields of spatial epidemiology and Machine Learning applications in public health.

Materials and methods

Data collection and study area

The retrospective cohort analytical study examined 554 sub-districts across four regions of Thailand (northern, central, northeastern, southern). We collected data from two main sources:

CCA case data

Information from four Population-Based Cancer Registries (PBCRs): northern (Lampang Cancer Hospital), central (Lop Buri Cancer Hospital), northeastern (Khon Kaen Provincial Cancer Registry), and southern region (Surat Thani Cancer Hospital) [23]. All the CCA cases were diagnosed between January 1, 2012, and December 31, 2021, based on the International Classification of Diseases for Oncology, 3rd Edition (ICD-O-3), with the specific codes: C22.1 (Intrahepatic bile duct), C24.0 (Extrahepatic bile duct), C24.8 (Overlapping lesion of biliary), and C24.9 (Biliary tract, NOS) (excluding C24.1, Ampulla of Vater) [24, 25]. Key variables included sex, age at diagnosis, birth date, ICD-O-3 code, address, and basis of diagnosis. Population data from the Office of the National Economic and Social Development Board [26] was used to calculate the age-standardized rates (ASR) by sex and age groups every five years between 2012 and 2021 (Table 1).

Table 1 Descript of dependent and independent variables used in the study

Variable Type	Variable Name	Description	Unit	Data Source
Dependent Variable	ASR of CCA	Incidence rate of CCA adjusted for age distribution using the Segi World standard population	Cases per 100,000 person-years	Calculated from cancer registry data using IACR guidelines
Independent Variables	Elevation	Average height above sea level in each sub-district	Meters	Department of Water Resources
	Distance from water sources	Average distance from population centers to nearest water body	Meters	Department of Water Resources
	Population density	Number of people per unit area	Persons/km^2	National Economic and Social Development Board
	Average rainfall	Mean annual precipitation	Millimeters/year	Thailand Meteorological Department
	Average temperature	Mean annual temperature	°C/year	Thailand Meteorological Department

Spatial variables

First, environment data (elevation, water source coordinates, and the size and extent of areas) from the Central Geoinformatics System and Services Project, Department of Water Resources, Ministry of Natural Resources and Environment [27]. Second, climatic data (average rainfall, average temperature, and coordinates for all meteorological stations) were obtained from the Thailand Meteorological Department using a statistical data request system [28]. All spatial variables were aggregated at the sub-district level (Table 1).

Study areas

The study covered four provinces representing the four main region of Thailand respective sizes and geographical coordinates (latitudes and longitudes): (i) Lampang province (Northern): 12,533.96 km^2, 17.2°−19.5°N, 98.9°−100.2°E; (ii) Lop Buri province (Central): 6,208.70 km^2, 14.6°−15.8°N, 100.3°−101.5°E; (iii) Khon Kaen province (Northeastern): 10,885.99 km^2, 15.6°−17.1°N, 101.6°−103.3°E; and (iv) Surat Thani province (Southern): 12,891.4 km^2, 8.3°−10.2°N, 98.5°−100.2°E, for each provinces representing the regions, respective [23].

Variables and measurement

ASR, Age-Standardized Rates; CCA, cholangiocarcinoma; IACR, International Association of Cancer Registries.

Statistical analysis

Incidence rate of CCA

The ASR was calculated for each sex and standardized using the Segi World standard population estimates [29]. The International Association of Cancer Registries (IACR) guidelines [30] were used to calculate the ASR of CCA cases in each sub-district.

Machine learning models

We implemented four different machine learning models to predict CCA incidence based on spatial variables. In our data management process, residential address codes were utilized as the key identifier for linking CCA cases data with all spatial factors. Prior to analysis, distribution testing was conducted for all variables. In cases where data exhibited abnormal distribution patterns (left or right skewness), variable transformation through logarithmic conversion was performed on all affected variables before proceeding with into machine learning models. Each model represents a different approach to predictive modeling, selected to provide a comprehensive comparison of techniques applicable to spatial epidemiology:

Linear regression

A statistical model that examines the linear relationship between the dependent variable (ASR of CCA) and multiple independent variables (spatial factors). We selected this model as a baseline comparison since it represents traditional statistical approaches and assumes linear relationships between variables.

Random Forest

An ensemble learning method that constructs multiple decision trees during training and outputs the average prediction of individual trees. Random Forests are well-suited for spatial epidemiology because they can capture non-linear relationships, handle interactions between variables, and are robust against overfitting. The algorithm works by bootstrapping samples of observations and variables to build diverse decision trees, with each tree contributing a vote toward the final prediction [31].

The Random Forest model was configured with the following specifications: number of trees = 500; variables randomly sampled at each split (mtry) = 2; minimum node size = 5; and Gini criterion was employed as the splitting criterion.

Neural Network

A computational model inspired by the human brain's neural structure, designed to recognize complex patterns through interconnected layers of nodes (neurons). Neural Networks process information through three main components: an input layer (receiving spatial variables), hidden layers (processing information through weighted connections), and an output layer (producing CCA incidence predictions). This architecture allows Neural Networks to model highly complex, non-linear relationships between spatial factors and disease incidence [32, 33]. The Neural Network utilized a $5 \rightarrow 15 \rightarrow 10 \rightarrow 1$ architecture with ReLU activation for hidden layers and linear activation for the output layer. Training employed a 0.01 learning rate with Adam optimizer, L2 regularization (weight decay = 0.0001), batch size of 32, and 200 epochs with early stopping to optimize performance.

Extreme Gradient Boosting (XGBoost)

An advanced implementation of gradient boosting that builds models sequentially, with each new model correcting errors made by previous ones. XGBoost has three key components: (i) a loss function to evaluate model accuracy, (ii) weak learners (typically decision trees) that perform slightly better than random guessing, and (iii) an additive model that combines weak learners into a strong predictive system. XGBoost includes regularization techniques to prevent overfitting, making it potentially valuable for spatial prediction with limited data [34]. The XGBoost model was configured with a learning rate of 0.05, maximum tree depth of 6, and minimum child weight of 3. Both subsample and column sample ratios were set at 0.8. For regularization purposes, alpha and lambda parameters were established at 0.2 and 0.1, respectively. The model was trained using 1000 boosting rounds with an early stopping mechanism to prevent overfitting and optimize performance.

Model training and validation

For model development and evaluation, we randomly split the dataset into training (70%) and testing (30%) subsets. This ratio was selected to balance the need for adequate training data while ensuring sufficient test data for reliable performance evaluation, given our sample size constraints. The 70:30 split is widely used in machine learning applications and provides a good compromise between these competing needs.

While we considered alternative splitting ratios (80:20, 90:10) by evaluating the same model as all real analyses, our preliminary analyses showed that the 70:30 split offered the optimal balance between model learning and validation for our dataset size. With approximately 554 sub-districts in our study, this split provided 388 sub-districts (4,465 cases) for training and 166 for testing (1,914 cases)—sufficient numbers for both robust model training and meaningful validation without risking overfitting.

Table 1 illustrates our complete research methodology from data collection through model evaluation. The process began with gathering CCA case data from four regional cancer registries and spatial data from government databases. After preprocessing, which included calculating ASR values and standardizing spatial variables, we implemented the 70:30 random split stratified by region to maintain proportional representation. Each model was trained using identical training data and hyperparameter optimization techniques, then evaluated on the common test set using RMSE, R^2, and visual assessment via scatter plots.

Model evaluations

We implemented a comprehensive evaluation framework using three complementary approaches to ensure robust assessment of model performance:

Root Mean Square Error (RMSE)

RMSE quantifies prediction errors in the same units as the dependent variable, giving greater weight to large errors—critical in health applications where significant errors can have serious consequences. This metric calculates the square root of the average squared differences between predicted and actual CCA incidence values:

$$RMSE = \sqrt{[\sum (predicted - actual)2/n]}$$

Lower RMSE values indicate better model performance with less prediction error. We selected RMSE over alternative metrics like Mean Absolute Error (MAE) because RMSE gives greater weight to large errors through the squaring mechanism, making it particularly valuable for health applications where large prediction errors could have significant consequences for resource allocation and intervention planning. This sensitivity to outliers helps identify models that might perform well on average but produce concerning errors in certain regions or incidence ranges.

R-squared (R^2)

This coefficient of determination measures the proportion of variance in the dependent variable (ASR of CCA) that is explained by the model's independent

variables (spatial factors). By providing a straightforward scale from 0 to 1, it allows us to measure the proportion of variance explained in CCA incidence and facilitates meaningful comparisons with previous research findings:

R^2 = *1 - (Sum of Squared Residuals/Total Sum of Squares)*

R^2 values range from 0 to 1, with values closer to 1 indicating that the model explains a greater proportion of the variance in CCA incidence, suggesting better predictive performance. For each model, we calculated 95% confidence intervals for R^2 values using bootstrap resampling with 1000 iterations to quantify the uncertainty in our performance estimates and allow for more rigorous statistical comparison between models.

Scatter plots

We created scatter plots to visualize the relationship between predicted and actual CCA incidence values for each model. These visual representations serve multiple analytical purposes:

- Identifying patterns in prediction accuracy across different incidence levels.
- Revealing potential systematic biases (such as consistent over-prediction in high-incidence areas).
- Detecting heteroscedasticity in prediction errors.
- Identifying regional clusters or outliers that might require special attention.

We enhanced these scatter plots with a 45-degree reference line representing perfect prediction, regression lines showing actual trends, and color-coding by region to enable deeper visual analysis of model performance.

After comprehensive comparison of these models, we conducted variable importance analysis using the best-performing model (Random Forest) to identify the key spatial predictors of CCA incidence. This analysis quantifies the mean decrease in prediction accuracy when each variable is excluded from the model while keeping all others constant. The approach involved:

1. Training the optimal Random Forest model on the complete dataset
2. Permuting each predictor variable one at a time (effectively removing its information while maintaining the same data structure)
3. Measuring the resulting decrease in prediction accuracy
4. Ranking variables by their impact on model performance

This permutation-based approach offers advantages over alternative variable importance methods as it directly measures the impact on the model's predictive performance rather than changes in node purity, providing more interpretable results that directly relate to our prediction goals.

For model implementation, we used the Random Forest, Nural Network, XGBoost, and stats packages in R. We performed all analyses and visualizations using R software version 4.2.1 (R Core Team) [35] with RStudio software version 1.4.1 [36]. Spatial data processing utilized the sf and raster packages, while visualization employed ggplot2 with custom themes for optimal clarity. Statistical validation, including confidence interval calculation, was implemented using the resamples.

Results

Demographic and spatial characteristics

From all the 6,379 CCA cases, most were males (4,075 cases; 63.9%) with a mean age of 66.2 years (standard deviation = 11.07 years) The northeastern region accounted for the majority of cases (3,898 cases; 61.1%), followed by the northern (1,695 cases; 26.6%), central (624 cases; 9.8%), and southern regions (162 cases; 2.5%), respectively (Table 2).

CCA incidence by region and sex

The overall ASR of CCA was 8.9 per 100,000 person-years (95% CI: 8.7 to 9.2), with males having a substantially higher incidence (12.5 per 100,000 person-years, 95% CI: 12.1 to 12.9) than females (5.9 per 100,000 person-years, 95% CI: 5.6 to 6.1). The northeastern region showed the highest incidence rates for both sexes (ASR = 13.4 per 100,000 person-years, 95% CI: 12.9 to 13.8), followed by the northern (ASR = 11.2 per 100,000 person-years, 95% CI: 10.6 to 11.7), central (ASR = 4.8 per 100,000 person-years, 95% CI: 4.5 to 5.2), and southern regions (ASR = 1.1 per 100,000 person-years, 95% CI: 0.9 to 1.3) (Table 3).

Table 4 and Fig. 1 (a-e) show the comparative performance of the four machine learning models across different regions. For the overall dataset, the Random Forest model demonstrated superior performance with the highest R^2 values in both training (72.07%) and testing datasets (71.66%), and the lowest RMSE values (training = 8.991, testing = 9.022). The XGBoost model showed the second-best performance overall (training R^2 = 70.57%, RMSE = 9.719 & testing R^2 = 68.30%, RMSE = 0.904), followed by Neural Network (training R^2 = 57.25%, RMSE = 11.044 & testing R^2 = 56.81%, RMSE = 11.076) and Linear Regression (training R^2 = 9.88%, RMSE = 16.034 & testing R^2 = 8.52%, RMSE = 16.078).

Table 2 Demographic characteristics of CCA in Thailand between 2012 and 2021

Characteristics	Number	Percentage
Sex		
Male	4,075	63.9
Female	2,304	36.1
Age of diagnosis (years)		
15–19	2	0.1
20–24	6	0.1
25–29	7	0.1
30–34	12	0.2
35–39	42	0.6
40–44	121	1.9
45–49	273	4.3
50–54	464	7.3
55–59	800	12.5
60–64	1,017	15.9
65–69	1,064	16.7
70–74	1,027	16.1
75 +	1,544	24.2
Mean (standard deviation)	66.2 (11.07)	
Median (Min: Max)	67.0 (19: 98)	
Elevation (meters)		
Mean (standard deviation)	187.4 (164.37)	
Median (Min: Max)	179 (5: 915)	
Distance from water sources (meters)		
Mean (standard deviation)	107.1 (1565.57)	
Median (Min: Max)	0 (0: 35,986.3)	
Population density (person/km^2)		
Mean (standard deviation)	213.1 (573.56)	
Median (Min: Max)	118.8 (3.1: 9251.2)	
Average rainfall (millimeter/year)		
Mean (standard deviation)	117.7 (31.24)	
Median (Min: Max)	109.4 (48.4: 240.7)	
Average temperature (°C/year)		
Mean (standard deviation)	28.4 (0.69)	
Median (Min: Max)	28.3 (26.9: 30.3)	
Regions		
Northern	1,695	26.6
Central	624	9.8
Northeastern	3,898	61.1
Southern	162	2.5

Model performance varied substantially across regions. In the northern region, all models achieved higher R^2 values than in other regions, with Random Forest showing the best performance (testing $R^2 = 87.30\%$). The central region showed moderate performance across models, with Random Forest again performing best (testing $R^2 = 77.17\%$). In the northeastern region, Random Forest maintained the highest performance (testing $R^2 = 76.81\%$). The southern region showed more variability in model performance, with XGBoost achieving the highest testing R^2 (63.04%) (Table 4).

The southern region displayed a distinct pattern, with XGBoost achieving the highest testing R^2 (63.04%), significantly outperforming Random Forest (41.08%). This region also showed the largest gaps between training and testing performance across all models, suggesting potential overfitting challenges in this region with the smallest sample size (Table 4, Fig. 1e).

Analysis of the scatter plots confirmed these quantitative findings, showing tighter clustering around the diagonal line of perfect prediction for Random Forest and XGBoost models, particularly in the northern and northeastern regions. Linear Regression consistently showed poor alignment with the diagonal, especially at higher ASR values, highlighting its inability to capture the non-linear relationships that characterize CCA's spatial epidemiology.

Variable importance analysis

The variable importance analysis of the Random Forest model identified elevation as the most important predictor, with a mean decrease in accuracy of 32.4% when removed from the model. Distance from water sources ranked second in importance, followed by population density and average temperature. Average rainfall showed the least influence on prediction accuracy (Fig. 2).

Discussion

The study presented compared the prediction performance of various Machine Learning approaches for spatial modeling of CCA incidence in Thailand. Our findings revealed complex patterns in model performance across different regions and identified key environmental determinants of CCA distribution.

The Random Forest model demonstrated superior overall predictive capability (testing $R^2 = 71.66\%$), consistently outperforming other approaches across most regions. This exceptional performance can be attributed to several advantages that make Random Forest particularly suitable for spatial epidemiological data: its ability to capture non-linear relationships, robustness to outliers, capacity to handle interactions between variables without explicit specification, and effective balance between model complexity and generalizability. In regions with larger sample sizes, Random Forest maintained minimal differences between training and testing performance, indicating excellent generalizability. These findings align with previous research by Tsilimigras et al. [16], who demonstrated that Random Forest models achieved 85% accuracy in predicting CCA phenotypes and patient

Table 3 Incidence of CCA by sex in each region of Thailand between 2012 and 2021

Region	Male		Female		Both sexes	Both sexes
	ASR	95% CI	ASR	95% CI	ASR	95% CI
Northern	14.7	13.8 to 15.6	7.9	7.3 to 8.6	11.2	10.6 to 11.7
Center	6.6	5.9 to 7.3	3.4	2.9 to 3.8	4.8	4.5 to 5.2
Northeastern	19.2	18.4 to 19.9	8.5	8.0 to 8.9	13.4	12.9 to 13.8
Southern	1.5	1.2 to 1.8	0.8	0.6 to 1.0	1.1	0.9 to 1.3
Total	12.5	12.1 to 12.9	5.9	5.6 to 6.1	8.9	8.7 to 9.2

ASR Age-Standardized Rates; *CI* Confidence Interval

Table 4 The Machine Learning models for predictions CCA in Thailand

Region/Model	Training		Testing	
	RMSE	R^2 (%)	RMSE	R^2 (%)
Overall				
Linear Regression	16.034	9.88	16.078	8.52
Random Forest	**8.991**	**72.07**	**9.022**	**71.66**
Neural Network	11.044	57.25	11.076	56.81
XGBoost	9.719	70.57	0.904	68.30
Northern				
Linear Regression	18.955	13.40	19.708	10.41
Random Forest	**7.011**	**88.19**	**7.422**	**87.30**
Neural Network	8.169	83.93	8.554	83.10
XGBoost	7.927	87.08	8.456	86.16
Central				
Linear Regression	4.278	8.20	4.326	7.05
Random Forest	**2.147**	**79.25**	**2.296**	**77.17**
Neural Network	3.105	51.67	3.190	49.60
XGBoost	2.074	83.28	2.350	77.52
Northeastern				
Linear Regression	12.896	10.36	13.380	10.06
Random Forest	**6.673**	**77.95**	**7.131**	**76.81**
Neural Network	9.613	50.19	10.127	48.52
XGBoost	7.360	76.24	7.852	75.14
Southern				
Linear Regression	2.358	5.87	2.477	3.77
Random Forest	1.209	76.65	1.884	41.08
Neural Network	1.563	56.72	1.664	53.59
XGBoost	**1.057**	**80.81**	**1.462**	**63.04**

RMSE Root Means Square Error; *R2 R-squared* Extreme Gradient Boosting

outcomes. Similarly, Liu et al. [37] found Random Forest achieved superior performance (AUC = 0.86) in spatial cancer risk prediction studies. Our results parallel to those of Thongpeth et al. [38], who compared various modeling approaches for healthcare predictions in Thailand and found Random Forest consistently outperformed other Machine Learning methods.

XGBoost showed strong performance overall (testing $R^2 = 68.30\%$) and performed exceptionally well in the southern region (testing $R^2 = 63.04\%$). However, it demonstrated greater inconsistency between training and testing performance, particularly in regions with smaller sample sizes. This pattern suggests potential overfitting issues—a known limitation of boosting methods when applied to smaller datasets. This finding differs from results reported by Wu et al. [39], who found XGBoost achieved the highest accuracy (AUC = 0.892) in predicting CCA outcome, and Chaudhary et al. [40], who reported XGBoost outperforming traditional methods with 89.2% accuracy. This discrepancy likely reflects differences between clinical prediction contexts (featuring individual-level variables) and our spatial analysis (using aggregated environmental factors at the sub-district level).

Neural Networks showed moderate but consistent performance across regions (overall testing $R^2 = 56.81\%$), with the smallest gap between training and testing metrics. This finding contrasts with Zhang et al. [41], who found Neural Networks achieved superior performance in cancer prediction. However, our results align with Wang et al.'s findings [42] that tree-based models often outperform Neural Networks in spatial disease prediction, particularly in environments with complex ecological interactions.

The striking performance gap between machine learning models and Linear Regression (testing $R^2 = 8.52\%$) confirms that CCA's spatial distribution follows complex, non-linear patterns that cannot be adequately captured by traditional statistical approaches. This finding has important methodological implications for future spatial epidemiology studies, suggesting that machine learning approaches should be preferred for similar complex spatial health phenomena. The dramatic performance differential stems from fundamental advances that machine learning brings to spatial epidemiology: superior ability to capture geographic variations in disease-environment relationships, better

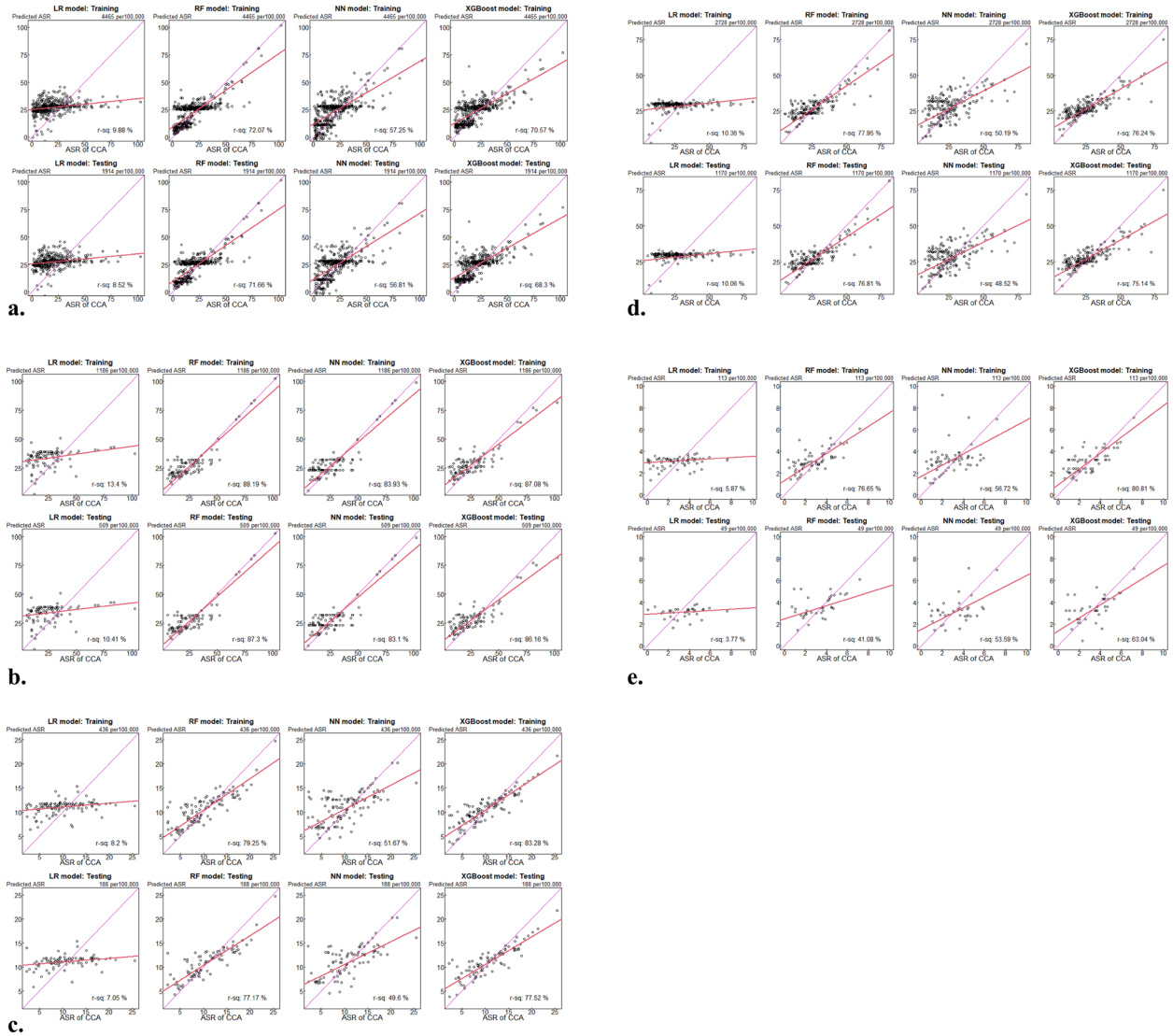

Fig. 1 Scatter plots comparing predicted versus observed CCA rates across Thailand by Machine Learning models. LR, Linear Regression; RF, Random Forest; NN, Neural Network; XGBoost, Extreme Gradient Boosting, **b** Scatter plots comparing predicted versus observed CCA rates in Northern Thailand by Machine Learning models. LR, Linear Regression; RF, Random Forest; NN, Neural Network; XGBoost, Extreme Gradient Boosting. **c** Scatter plots comparing predicted versus observed CCA rates in Central Thailand by Machine Learning models. LR, Linear Regression; RF, Random Forest; NN, Neural Network; XGBoost, Extreme Gradient Boosting. **d** Scatter plots comparing predicted versus observed CCA rates in Northeastern Thailand by Machine Learning models. LR, Linear Regression; RF, Random Forest; NN, Neural Network; XGBoost, Extreme Gradient Boosting. **e** Scatter plots comparing predicted versus observed CCA rates in Southern Thailand by Machine Learning models. LR, Linear Regression; RF, Random Forest; NN, Neural Network; XGBoost, Extreme Gradient Boosting

handling of spatial dependencies that violate independence assumptions in traditional models, and effective integration of multi-scale spatial interactions without requiring explicit hierarchical modeling.

Our analysis revealed substantial regional variations in both CCA incidence and model performance. All models performed best in the northern region (testing R^2 up to 87.30%), followed by northeastern (76.81%) and central regions (77.52%), with more modest performance in the southern region (63.04%). These patterns align with findings from Kaewpitoon et al. [13], who observed varying prediction accuracies across different Thai regions using GIS-based analysis. The exceptional performance in the northern region suggests that environmental factors strongly and consistently influence CCA risk in this area. By contrast, the more moderate performance in the southern region, despite using identical predictors, indicates that different etiological factors may be at play

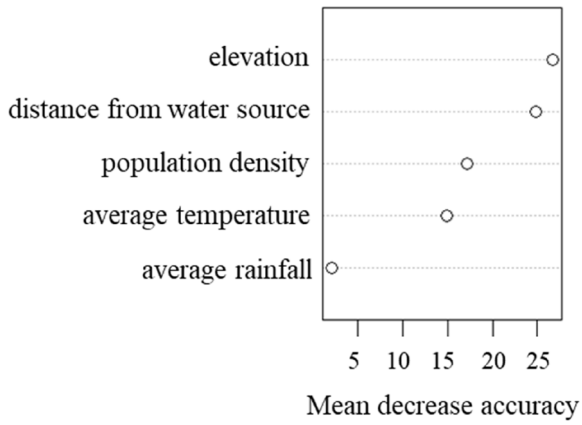

Fig. 2 Variable importance analysis showing relative impact of spatial predictors on CCA incidence

or that environmental relationships are more complex in this region. This regional heterogeneity in model performance highlights the importance of region-specific approaches to both disease modeling and public health intervention.

Our variable importance analysis identified elevation as the most significant predictor of CCA incidence, followed by population density, distance from water sources, and average rainfall. Elevation likely serves as a proxy for multiple ecological factors: it influences water flow patterns and drainage characteristics critical for *O. viverrini*'s lifecycle, affects agricultural practices (particularly rice cultivation associated with increased human-water contact), and historically shaped settlement patterns in ways that overlap with endemic zones. The importance of water-related variables aligns with the established understanding of *O. viverrini*'s lifecycle, which requires freshwater environments for transmission through intermediate snail hosts and fish. These environmental determinants help explain the pronounced regional disparities in CCA incidence observed in our study. The northeastern region's high rates (13.4 per 100,000 person-years) [43] correspond with previous studies documenting high *O. viverrini* prevalence in this area, while the southern region's low rates (1.1 per 100,000 person-years) reflect minimal *O. viverrini* endemicity. The topographical and hydrological conditions of northeastern Thailand—characterized by low-elevation plateaus with numerous water bodies—create ideal conditions for parasite transmission, which our models effectively captured through environmental predictors [2].The implementation of Machine Learning for spatial epidemiology of cancer aligns with similar initiatives in other countries. Qiao et al. [44] successfully implemented Machine Learning models for cancer prediction in China, achieving accuracy rates of 87.5% using ensemble methods similar to our Random Forest approach. Similarly, Kim et al. [45] demonstrated the effectiveness of spatial Machine Learning in predicting cancer patterns in South Korea, with Random Forest models showing high accuracy (AUC 0.89). The comparable performance of our models (particularly in the northern region with $R^2 = 87.30\%$) suggests that our methodological approach represents current international best practices in spatial health modeling. However, our southern region results (maximum $R^2 = 63.04\%$) highlight an important limitation: machine learning approaches remain sensitive to sample size constraints. The southern region's substantially lower CCA incidence resulted in fewer cases for model training, potentially limiting predictive accuracy.

The ability to predict CCA incidence with high accuracy has significant implications for public health planning in Thailand. Our findings can transform CCA control efforts in several crucial ways. Rather than implementing uniform screening programs, health authorities can use our predictive models to identify high-risk communities based on environmental factors. This approach would enable more efficient allocation of screening resources to areas with the highest predicted CCA risk, potentially improving early detection rates in a cost-effective manner. Understanding the relationship between environmental factors and CCA risk can guide targeted interventions addressing specific risk factors. For example, communities in low-elevation areas near water sources might benefit from enhanced water treatment initiatives, while education programs about proper fish cooking practices could be prioritized in areas with high predicted risk. The varying performance of models across regions suggests that a"one-size-fits-all"approach may not be optimal. In the northeastern and northern regions, where environmental factors strongly predict CCA risk, targeted interventions based on spatial risk factors are likely to be effective. The southern region, with its distinct epidemiological profile, may require different approaches.

Given the importance of elevation and water-related variables, climate change could potentially alter CCA risk patterns through changes in precipitation, temperature, and water body characteristics. Rising temperatures and changing precipitation patterns could shift the geographic distribution of suitable habitats for *O. viverrini*'s intermediate hosts. Our predictive framework provides a baseline for modeling future scenarios under different climate projections. The translation of our findings into practical public health applications aligns with Thailand's National Artificial Intelligence Strategy (NAIS) Action Plan for 2022–2027 [22]. Our machine learning approach to disease prediction represents a concrete

implementation of Strategy 4, which focuses on developing intelligent technologies to address national challenges. By demonstrating the superior performance of advanced analytical methods over traditional approaches, we provide evidence-based support for broader adoption of AI-driven approaches in public health planning across Thailand.

Our study features several methodological strengths that enhance the reliability and applicability of our findings. The comprehensive inclusion of 6,379 CCA cases from four population-based cancer registries provides a robust epidemiological foundation rarely achieved in spatial modeling studies. Our comparative evaluation of multiple machine learning approaches offers methodological insights beyond single-model studies, while the variable importance analysis provides a quantitative hierarchy of environmental determinants that advances epidemiological understanding beyond traditional association studies. Nevertheless, several limitations warrant acknowledgment. Despite our large overall sample, the regional distribution was uneven, with relatively few cases in the southern region (162 cases; 2.5%). This imbalance likely contributed to the lower model performance in that region and highlights a common challenge in modeling rare diseases across heterogeneous geographies. Our analysis relied on spatial variables available at the sub-district level, potentially missing finer-scale variations that could influence local CCA risk patterns. While our models effectively captured spatial patterns, they did not incorporate temporal dynamics of CCA development—a significant consideration given the often decades-long lag between *O. viverrini* exposure and cancer development. Future research should address these limitations through several approaches. Incorporating village-level socioeconomic indicators, local food consumption patterns, and sanitation infrastructure data could enhance prediction accuracy by capturing behavioral determinants of *O. viverrini* exposure. Studies by Songserm et al. [6] suggest these factors might explain 15–20% of the variance currently not captured by environmental variables alone. Developing models that incorporate both spatial patterns and temporal trends could provide insights into how CCA incidence evolves over time in response to environmental changes and public health interventions. Using more detailed environmental data, including high-resolution remote sensing of surface water characteristics and land use patterns, could improve prediction accuracy by better capturing habitat suitability for intermediate hosts. Building on our identified environmental predictors, future research should model how climate change might alter CCA risk distribution through changes in temperature, precipitation, and hydrological patterns. The methodological advances demonstrated in our study—particularly the superior performance of machine learning approaches compared to traditional statistical methods—should inform future spatial epidemiology research for other environmentally-mediated diseases in Thailand and beyond.

Conclusions

The incidence of CCA in Thailand presented in this study, found that most of the CCA cases occur in the Northeastern, Northern, Central, Southern region, respectively. In analyzing predictive models for CCA incidence in Thailand using R^2 and RMSE, the Random Forest model has emerged as the most effective approach with 71.66% prediction, followed by the XGBoost model (68.30% prediction), and the Neural Network model (56.81% prediction), respectively. In each region different Machine Learning models with regional variations highlighted the complexity of cholangiocarcinoma epidemiology across different parts of Thailand. Furthermore, spatial factors demonstrated the predictive capabilities for ASR of CCA. This national finding has pioneered the CCA distribution in Thailand and has developed a spatial-based approach to support disease control. The research presented in this paper has pointed to opportunities for examining additional geographical variables in future studies.

Acknowledgements

The authors thank all members of the four Population-based Cancer Registries, namely Lampang Cancer Hospital, Lop Buri Cancer Hospital, Khon Kean provincial cancer registry, Srinagarind Hospital, Faculty of Medicine, Khon Kaen University, Surat Thani Cancer Hospital. Ulster University is acknowledged for supporting the writing of this research paper during Oraya Sahat's visit to Belfast Campus from January 12 to February 12, 2025.

Authors' contributions

SK1 served as the corresponding author of this study. OS was the principal author and collaborated with all authors to develop the study conception and design. AL conducted the data analysis with support from SK1 and SK2. MGC reviewed the methodology and results along with IE reviewed and improved the language of the article. All authors reviewed previous versions of the manuscript and approved the final version.

Funding

Not applicable.

Data availability

The population-based cancer registry data analyzed in this study, which includes confidential personal information from four Thai regions, cannot be shared publicly due to privacy regulations. The data are available from the corresponding author upon reasonable request from interested researchers.

Declarations

Ethics approval and consent to participate

This study utilized secondary data from four PBCRs, which did not involve the collection of individuals' identifying information. Therefore, individual informed consent was not required. This study received ethical approval from the Human Research Ethics Committees of all four data sources: Lampang Cancer Hospital (No. 10/2567), Lop Buri Cancer Hospital (No. LEC 6647), Khon Kaen University, where the consideration of human research ethics is in accordance with the Helsinki Declaration (No. HE671027), and Surat Thani Cancer Hospital (No. SCH_EC_01/2567).

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]Student of Doctor of Public Health Program, Faculty of Public Health, Khon Kaen University, Khon Kaen, Thailand. [2]Department of Epidemiology and Biostatistics, Faculty of Public Health, Khon Kaen University, Khon Kaen, Thailand. [3]Department of Mathematics and Computer Science, Faculty of Science and Technology, Prince of Songkla University, Pattani Campus, Pattani, Thailand. [4]School of Computing, Ulster University, Northern Ireland, Belfast Campus, Belfast BT15 1 AP, UK.

Received: 11 February 2025 Accepted: 9 May 2025
Published online: 07 June 2025

References

1. Banales JM, Marin JJG, Lamarca A, Rodrigues PM, Khan SA, Roberts LR, et al. Cholangiocarcinoma 2020: the next horizon in mechanisms and management. Nat Rev Gastroenterol Hepatol. 2020;17(9):557–88. https://doi.org/10.1038/s41575-020-0310-z.
2. Sriamporn S, Pisani P, Pipitgool V, Suwanrungruang K, Kamsa-ard S, Parkin DM. Prevalence of Opisthorchis viverrini infection and incidence of cholangiocarcinoma in Khon Kaen. Northeast Thailand Trop Med Int Health. 2004;9(5):588–94. https://doi.org/10.1111/j.1365-3156.2004.01234.x.
3. Kamsa-Ard S, Luvira V, Pugkhem A, Luvira V, Thinkhamrop B, Suwanrungruang K, et al. Association between praziquantel treatment and cholangiocarcinoma: a hospital-based matched case-control study. BMC Cancer. 2015;15:776. https://doi.org/10.1186/s12885-015-1788-6.
4. Shin H-R, Oh J-K, Masuyer E, Curado M-P, Bouvard V, Fang Y-Y, et al. Epidemiology of cholangiocarcinoma: an update focusing on risk factors. Cancer Sci. 2010;101:579–85. https://doi.org/10.1111/j.1349-7006.2009.01458.x.
5. Honjo S, Srivatanakul P, Sriplung H, Kikukawa H, Hanai S, Uchida K, et al. Genetic and environmental determinants of risk for cholangiocarcinoma via Opisthorchis viverrini in a densely infested area in Nakhon Phanom, northeast Thailand. Int J Cancer. 2005;117(5):854–60. https://doi.org/10.1002/ijc.21146.
6. Songserm N, Promthet S, Sithithaworn P, Pientong C, Ekalaksananan T, Chopjitt P, et al. Risk factors for cholangiocarcinoma in high-risk area of Thailand: role of lifestyle, diet and methylenetetrahydrofolate reductase polymorphisms. Cancer Epidemiol. 2012;36(2):e89-94. https://doi.org/10.1016/j.canep.2011.11.007.
7. Sripa B, Pairojkul C. Cholangiocarcinoma: lessons from Thailand. Curr Opin Gastroenterol. 2008;24:349–56. https://doi.org/10.1097/MOG.0b013e3282fbf9b3.
8. Sripa B, Kaewkes S, Sithithaworn P, Mairiang E, Laha T, Smout M, et al. Liver Fluke Induces Cholangiocarcinoma. PLoS Med. 2007;4(7): e201. https://doi.org/10.1371/journal.pmed.0040201.
9. Promthet S, Kamsa-Ard S, Vatanasapt P, Wiangnon S, Suwanrungruang K, Poomphakwaen K. Risk factors for cancers: a cohort study in Khon Kaen, Northeast Thailand. Office of the Health Promotion Foundation under the health information system development plan. 2010.
10. Kirby RS, Delmelle E, Eberth JM. Advances in spatial epidemiology and geographic information systems. Ann Epidemiol. 2017;27(1):1–9. https://doi.org/10.1016/j.annepidem.2016.12.001.
11. Mungthanee T. The Thai-Laos Culture and The Solution of Liver Fluke and Cholangiocarcinoma: A Case Study in The Middle Songkhram River Basin. AJHSS BUU. 2019;27(55):60–82.
12. Tamngam P, Pamulila S, Sarakum N, Inpang S. Knowledge, Attitude, and Consumption Behavior Associated with Cholangiocarcinoma in a Sub-District, Warinchamrab District, Ubon Ratchathani Province. J Sci Tech UBU. 2019;21(3):74–85.
13. Kaewpitoon SJ, Rujirakul R, Joosiri A, Jantakate S, Sangkudloa A, Kaewthani S, et al. GIS Database and Google Map of the Population at Risk of Cholangiocarcinoma in Mueang Yang District, Nakhon Ratchasima Province of Thailand. Asian Pac J Cancer Prev. 2016;17(3):1293–7. https://doi.org/10.7314/apjcp.2016.17.3.1293.
14. Blangiardo M, Cameletti M, Baio G, Rue H. Spatial and spatio-temporal models with R-INLA. Spat Spatiotemporal Epidemiol. 2013;7:39–55. https://doi.org/10.1016/j.sste.2013.07.003.
15. Haghbin H, Aziz M. Artificial intelligence and cholangiocarcinoma: Updates and prospects. World J Clin Oncol. 2022;13(2):125–34. https://doi.org/10.5306/wjco.v13.i2.125.
16. Tsilimigras DI, Hyer JM, Paredes AZ, Diaz A, Moris D, Guglielmi A, et al. A novel classification of intrahepatic cholangiocarcinoma phenotypes using machine learning techniques: an international multi-institutional analysis. Ann Surg Oncol. 2020;27(13):5224–32. https://doi.org/10.1245/s10434-020-08646-9.
17. Tsilimigras DI, Mehta R, Pawlik TM. ASO author reflections: use of machine learning to identify patients with intrahepatic cholangiocarcinoma who could benefit more from neoadjuvant therapies. Ann Surg Oncol. 2020;27(4):1120–1. https://doi.org/10.1245/s10434-020-08263-6.
18. Wiens J, Shenoy ES. Machine learning for healthcare: on the verge of a major shift in healthcare epidemiology. Clin Infect Dis. 2018;66(1):149–53. https://doi.org/10.1093/cid/cix731.
19. Grebovic M, Filipovic L, Katnic I, Vukotic M, Popovic T. Machine learning models for statistical analysis. Int Arab J Inf Technol. 2023;20(3A):505–14. https://doi.org/10.34028/iajit/20/3A/17.
20. Dhillon SK, Ganggayah MD, Sinnadurai S, Lio P, Taib NA. Theory and practice of integrating machine learning and conventional statistics in medical data analysis. Diagnostics. 2022;12(10): 2526. https://doi.org/10.3390/diagnostics12102526.
21. Saha S, Moorthi S, Wu X, Wang J, Nadiga S, Tripp P, et al. The NCEP climate forecast system version 2. J Clim. 2014;27(6):2185–208. https://doi.org/10.1175/JCLI-D-12-00823.1.
22. Ministry of Higher Education, Science, Research and Innovation and Ministry of Digital Economy and Society. the National Artificial Intelligence Action Plan with a vision for the development of Thailand 2022 – 2027 [Online] 2022. Available from: https://ai.in.th/wp-content/uploads/2022/12/20220726-AI.pdf. [inThai]. Cited 2024 Sep 12.
23. Rojanamatin J, Ukranum W, Supaattagorn P, Chiawiriyabunya I, Wongsena M, Chaiwerawattana A, et al. Cancaer in Thailand Volume. X, 2016–2018. Bankok: Medical Record and Databased Cancer Unit; 2021. [in Thai]
24. WHO. WHO | International Classification of Diseases for Oncology, 3rd Edition (ICD-O-3). 2013. Available from: http://www.who.int/classifications/icd/adaptations/oncology/en/. Cited 2023 Aug 15.
25. World Health Organization. International classification of diseases for oncology (ICD-O), 3rd ed., 1st revision. 2013. Available from: https://www.who.int/standards/classifications/other-classifications/international-classification-of-diseases-for-oncology. Cited 2023 Aug 21.
26. Office of the National Economic and Social Development Board. Population Projections for Thailand 2010–2040. Bangkok, Thailand: Office of the National Economic and Social Development Board; 2013.
27. Department of Water Resources Ministry of Natural Resources and Environment. Central Geo-Informatics System and Services Project. 2022. Available from: https://webgis.dwr.go.th/. Cited 2023 Sep 9.
28. Information Technology Center Meteorological Department. Meteorological statistics data request submission system. 2023. Available from: https://data-service.tmd.go.th/index.php. Cited 2023 Sep 13.
29. Bray F, Colombet M, Aitken JF, Bardot A, Eser S, Galceran J, et al. Cancer Incidence in Five Continents, Vol. XII (IARC CancerBase No. 19). Lyon: International Agency for Research on Cancer. 2023. Available from: https://ci5.iarc.who.int. Cited 2024 Feb 28.
30. Boyle P, Parkin DM. Cancer registration: principles and methods. Statistical methods for registries. IARC Sci Publ. 1991;95:126–58.
31. Breiman L. Random forests. Mach Learn. 2001;45:5–32. https://doi.org/10.1023/A:1010933404324.
32. Krose B, Smagt PVD. An introduction to neural networks: The University of Amsterdam; 1996.
33. McCulloch WS, Pitts W. A logical calculus of the ideas immanent in nervous activity. Bull Math Biophys. 1943;5:115–33. https://doi.org/10.1007/BF02478259.
34. Chen T, Guestrin C, editors. Xgboost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining; 2016. https://doi.org/10.1145/2939672.2939785.

35. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria; 2023. Available from https://www.R-project.org/. Cited 2024 Jan 12.
36. Posit team. RStudio: Integrated Development Environment for R. Posit Software. Boston, PBC; 2023. Available from: http://www.posit.co/. Cited 2024 Jan 12.
37. Liu Y, Wu J, Liu M, Xu K, Guo Y, Wu J. Spatial epidemiology and machine learning methods for risk assessment of digestive tract cancers. Int J Environ Res Public Health. 2020;17(11): 3828. https://doi.org/10.3390/ijerph17113828.
38. Thongpeth W, Lim A, Wongpairin A, Thongpeth T, Chaimontree S. Comparison of linear, penalized linear and machine learning models predicting hospital visit costs from chronic disease in Thailand. Inform Med Unlocked. 2021;26: 100769.
39. Wu L, Zhou B, Yan C, Li M, Liu T, Zhu Q, et al. A deep learning model to predict survival outcomes in intrahepatic cholangiocarcinoma using histopathological images. Theranostics. 2021;11(15):7537–50. https://doi.org/10.7150/thno.59879.
40. Chaudhary K, Poirion OB, Lu L, Garmire LX. Deep learning-based multi-omics integration robustly predicts survival in liver cancer. Clin Cancer Res. 2018;24(6):1248–59. https://doi.org/10.1158/1078-0432.CCR-17-0853.
41. Zhang R, Xu J, Wang Y, Lu H, Miao Z, Han Z. Development and validation of a machine learning model for predicting illness trajectory and hospital resource utilization of patients with cholangiocarcinoma. JMIR Med Inform. 2021;9(4): e26586. https://doi.org/10.2196/26586.
42. Wang S, Yang DM, Rong R, Zhan X, Xiao G. Pathology image analysis using segmentation deep learning algorithms. Am J Pathol. 2019;189(9):1686–98. https://doi.org/10.1016/j.ajpath.2019.05.007.
43. Khuntikeo N, Loilome W, Thinkhamrop B, Chamadol N, Yongvanit P. A comprehensive public health conceptual framework and strategy to effectively combat Cholangiocarcinoma in Thailand. PLoS Negl Trop Dis. 2016;10(1): e0004293. https://doi.org/10.1371/journal.pntd.0004293.
44. Qiao Z, Sun N, Li X, Xia E, Li S, Chang Y. Using machine learning approaches for emergency room visit prediction based on electronic health record data. Stud Health Technol Inform. 2018;247:111–5. https://doi.org/10.3233/978-1-61499-852-5-111.
45. Kim BJ, Kim JH, Kim HS, Choi YH. Machine learning application for prediction of cholangiocarcinoma in patients with primary sclerosing cholangitis. J Hepatol. 2021;74(3):567–74. https://doi.org/10.1016/j.jhep.2020.10.038.

Publisher's Note