STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **x**, Month 202x, pp 0–13.
Published online in International Academic Press (www.IAPress.org)

Comparative Machine Learning Framework for Classifying Cancer Incidence in Iraq (2018–2020): Evidence from Random Forest, XGBoost, SVM, and Shrinkage-QDA

Omar Fawzi Salih Al-Rawi*

Department of Statistics and Informatics Techniques, Technical college of Management–Mosul, Northern Technical University, Mosul 41000, Iraq

Abstract Although the use of machine learning in cancer research is increasing, most studies conducted in Iraq have focused primarily on clinical diagnosis, medical image classification, and survival analysis, with little attention given to the classification of cancer incidence rates at the governorate level, based on unified national datasets. In addition, there is a lack of a comprehensive comparative framework for evaluating various machine learning algorithms under the same analytical conditions. To fill this gap, this study proposes a comparative machine learning framework for the classification of cancer incidence patterns in Iraqi governorates, using the 2018 and 2020 datasets. The analysis is focused on three age-specific cancer incidence variables, which are classified into low, medium, and high levels using a tertile-based classification method. Demographic and health-related predictors were incorporated and further enhanced by spatial feature engineering, including standardized variables, spatial lag effects, and spatial differences, to capture the spatial dependencies between neighboring governorates. The four classification algorithms applied and compared were Shrinkage Quadratic Discriminant Analysis (Shrinkage-QDA), Support Vector Machine (SVM), Random Forest, and Extreme Gradient Boosting (XGBoost). The performance of the models was assessed using accuracy and misclassification rates within a leave-one-out cross-validation (LOOCV) framework, which is suitable for small-sample spatial data. The results indicate a substantial variation in predictive performance across models and variables. Tree-based methods, particularly XGBoost and Random Forest, consistently demonstrated superior classification accuracy and lower misclassification rates, suggesting a greater capacity to model nonlinear relationships and spatial heterogeneity. On the other hand, SVM performed worst, and Shrinkage-QDA yielded moderate results. Temporal analysis showed a decrease in classification performance for one incidence category and an increase for the other categories in 2020. The overall results emphasize the importance of spatial feature engineering combined with machine learning, and provide evidence-based insights for cancer surveillance, spatial epidemiology, and public health planning in Iraq.

Keywords Machine Learning, Cancer Incidence, Spatial Analysis, Spatial Feature Engineering, Misclassification Rate

DOI: 10.19139/soic-2310-5070-3795

1. Introduction

The amalgamation of machine learning and spatial analysis has greatly assisted spatial epidemiology to a quite appreciable extent in better modelling of complex diseases and revealing underlying spatial patterns in data. Conventional statistical analyses often rest on linear assumptions. These assumptions can overlook nonlinear interactions and spatial dependencies in epidemiological data. On the contrary, machine learning modelling has flexible frameworks that are capable of establishing complex relations and posited that predictive performance and interpretability may be improved in spatial health studies [10, 13].

*Correspondence to: Omar Fawzi Salih Al-Rawi, (Email: omarfs@ntu.edu.iq). Department of Statistics and Informatics Techniques, Technical college of Management–Mosul, Northern Technical University, Mosul 41000, Iraq.

ISSN 2310-5070 (online) ISSN 2311-004X (print)

Many regions are worried about the variation of childhood cancer by geography and time. The factors affecting health indicators include exposure, socioeconomic and demographic structures, and their interactions. Ensamble learning models using Random Forest and Extreme Gradient Boosting (XGBoost) captures spatial heterogeneity of disease incidence effectively. The random forest offers a sufficient and stable performance due to bagging, while XGboost delivers a good prediction accuracy owing to gradient boosting and regularization techniques [1, 15].
Machine learning (ML) models' performance is dependent on how the response variable is related to the covariates. Spatial lag features are often used to capture spillover effects in space by pulling/mixing information from nearby space or regions. Spatial difference features can also be used to obtain local region discrepancies. This allows the detection of areas which are significantly different from that of their respective surrounding area [18, 12].
Additionally, a recent investigation reveals the importance of incorporating spatial-temporal dimensions into machine learning. Techniques with both space and time orientations enhance the understanding of the movement of diseases and improve predictive performance through spatial interactions and temporal evolution [5, 14].
Despite numerous major advancements, limitations still exist in the literature. Many studies look at spatial statistical models and ML models as separate modelling paradigms, not actively bringing spatial feature engineering into a predictive framework. In addition, a lot of past studies focus on continuous outcome modelling and the classification-based is limited. Using discretization techniques, we can also convert continuous variables into class attributes. For instance, one can use tertile-based classification for continuous variables. They will thus be transformed into class attributes for the BDA which use multiclass classification. This will preserve continuous variables variations as well [4, 8, 9].
Also, it is rarely or not evaluated temporal validation. It is essential to assess if the model's performance remains consistent over time.
In light of these gaps, the current research proposes a hybrid spatial classification framework where machine learning models and spatial feature engineering techniques can be employed to analyze childhood cancer incidence in Iraqi governorates. When we take original variables along with spatial lag and spatial difference features, the proposed Framework captures spatial dependence, local heterogeneity and Nonlinear Constraints. An integrated approach that bridges the literature gap between spatial econometrics and machine learning is offered, together with a comprehensive framework for spatial disease analysis and evidence-based public health decision-making.

2. Theoretical Framework

2.1. Construction of Dependent Variables

The three dependent variables considered in the research displayed the spatial distribution of the incidence of cancer of children in Iraqi governorates according to age groups, such as children under five years (x3), children that are five to less than 10 years (x4), and children that are 10 to less than 15 years (x5). All the dependent variables are analyzed separately in a different classification framework to assess age-specific spatial heterogeneity, which typifies most geographically referenced health data [2].
To enable classification modeling, we transform each continuous dependent variable into a three-level categorical variable (Low, Medium, High) using a tertile (quantile-based) classification scheme. There are three groups in the approach, All the groups are of equal size. Order of observations are preserved in any case. Further, the structure is easier to model. [17] Let (y) be the observed value to the corresponding threshold for the region as:

$$q_1 = Q\left(\frac{1}{3}\right), q_2 = Q\left(\frac{2}{3}\right) \tag{1}$$

The defined categorized variable $y_i$is.:

$$Y_i = \begin{cases} \text{Low}, & y_i \leq q_1 \\ \text{Medium}, & q_1 < y_i \leq q_2 \\ \text{High}, & y_i > q_2 \end{cases} \tag{2}$$

Using this discretization approach allows multi-class classification algorithms to be applied while preserving important spatial variation in the data. In addition, it allows the performance of models to be compared consistently across age groups in an epidemiology analysis [8, 9].

2.2. Independent Variables and Standardization

The text illustrates the classification of cancer cases and the inclusion of stomach ache as a variable.
In order to ensure the comparability of all variables and the unlikeliness that the scale of the variables may have affected the estimated model, z-score normalization was performed on all the model's predictors:

$$z_{ij} = \frac{x_{ij} - x_j}{s_j} \tag{3}$$

The meaning of x_{ij} is value of j variable of i region, x_j is mean and s_j is standard deviation of j variable. Machine learning algorithms such as SVM and distance-based methods are sensitive to level differences across a set of features often crucial changes [9].

2.3. Spatial Feature Engineering

To capture spatial dependence and heterogeneity as explicitly as possible, the present study organizes three categories of features; derive standardized variables, spatial lag variables and spatial difference variables.

2.3.1. Spatial Lag Variables : The weighted average of the same variable from neighboring observations is known as a spatial lag variable. If $W = w_{ij}$ is a spatial weights matrix then spatial lag of variable x_j is given by.:

$$L_j = Wx_j \tag{4}$$

or equivalently for region i:

$$L_{ij} = \sum_{k=1}^{n} w_{ik}x_{kj} \tag{5}$$

It means the spatial units are not independent and what happens in one region also impacts what happens in a neighbouring region [2]. The model introduces spatial lag variables to account for spatial autocorrelation and the interaction effects of the regions.

2.3.2. Spatial Difference Variables The main variables used to distinguish regions in space are designed to measure how far the region is from neighbouring regions. The definitions of these variables are.

$$D_j = x_j - Wx_j \tag{6}$$

or:

$$D_{ij} = x_{ij} - \sum_{k=1}^{n} w_{ik}x_{kj} \tag{7}$$

The features successfully capture spatial heterogeneity and localize regions where spatial context is significant. It is especially useful in spotting irregularities in space and localized risk patterns.

2.3.3. Final Feature Representation : The final feature matrix consists of all these components:

$$X^* = [X, WX, X - WX] \tag{8}$$

By using this full representation, the model can account for intrinsic characteristics, spatial dependence as well as local deviation simultaneously.

2.4. Classification Models

The utilized classification models were Shrinkage-QDA, SVM, Random Forest, and XGBoost as all four models can estimate both linear and non-linear relations (Patterson et al. 2018).

2.4.1. Shrinkage Quadratic Discriminant Analysis (Shrinkage-QDA) : According to QDA, the classes are multivariate normal with class-specific covariance matrices. A discriminant function's definition:

$$\delta_k(x) = -\frac{1}{2}\ln\Sigma_k - \frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k) + \ln(\pi_k) \tag{9}$$

A shrinkage estimator is employed to mitigate irregularities in the covariance estimation process.

$$\Sigma_k^{(\lambda)} = (1-\lambda)\Sigma_k + \lambda\,\mathrm{diag}(\Sigma_k) \tag{10}$$

This regularization improves robustness and reduces overfitting. The Shrinkage-QDA method enhances the stability and accuracy of classification by regularizing covariance matrix estimation, thereby reducing overfitting and improving performance in high-dimensional settings. It is commonly applied in fields such as bioinformatics, financial modeling, medical diagnostics, and image recognition, where datasets often contain a large number of features with limited observations [7].

2.4.2. Support Vector Machine (SVM) : SVM is the method to build an optimal separating hyperplane using optimization problem:

$$\min_{w,b,\xi}\frac{1}{2}\parallel w\parallel^2 + C\sum_{i=1}^{n}\xi_i \tag{11}$$

subject to:

$$y_i\left(w^T\phi(x_i)+b\right) \geq 1-\xi_i \tag{12}$$

Utilizing the kernel of the radial basis function:

$$K(x_i,x_j) = \exp\left(-\gamma\parallel x_i - x_j\parallel^2\right) \tag{13}$$

SVM effectively models nonlinear decision boundaries, SVM provides strong generalization performance by maximizing the margin between classes while controlling model complexity, making it robust to overfitting, especially in high-dimensional spaces. It is widely applied in text classification, image recognition, bioinformatics, and handwriting recognition, where nonlinear relationships and complex data structures are present [6].

2.4.3. Random Forest : Random forest is an ensemble machine learning algorithm that builds upon decision trees.

$$\hat{Y}(x) = \mathrm{mode}\{T_1(x),\ldots,T_B(x)\} \tag{14}$$

By combining bootstrap sampling and random feature selection, it reduces variance and improves generalization performance,random Forest enhances predictive accuracy and robustness by reducing variance through ensemble averaging, making it less prone to overfitting compared to individual decision trees. It is widely applied in classification and regression tasks such as medical diagnosis, financial prediction, fraud detection, and environmental modeling, particularly when dealing with complex and high-dimensional datasets [3].

2.4.4. XGBoost : XGBoost builds a decision tree additive model:

$$\hat{y}_i = \sum_{m=1}^{M} f_m(x_i) \tag{15}$$

By lowering the subsequent objective function:

$$\mathcal{L} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{m=1}^{M} \Omega(f_m) \tag{16}$$

with regularization term:

$$\Omega\left(f\right) = \gamma T + \frac{1}{2}\lambda \sum w_j^2 \tag{17}$$

This framework enhances both predictive accuracy and model regularization, XGBoost improves model performance by combining gradient boosting with regularization, leading to high accuracy, efficiency, and resistance to overfitting. It is widely applied in areas such as data mining competitions, financial forecasting, healthcare analytics, and classification problems involving large-scale and structured datasets [4].

2.5. Model Evaluation

The performance of the proposed models was evaluated using several evaluation metrics, which are robust to class imbalance.

2.5.1. Accuracy : Accuracy is calculated by taking the number of observations which are classified correctly, and dividing it by the overall number of observations. It can be represented as:

$$\text{Accuracy} = \frac{1}{n} \sum_{i=1}^{n} I(\hat{y}_i = y_i) \tag{18}$$

As per the above statement, y_i represents the predicted class label and y_i is defined as the true class label while I(.) is defined as the indicator function which is (1) when prediction is correct and is (0) otherwise. Accuracy has a very simple and easy to use nature. Nonetheless, it might be the most misleading classifier for imbalanced classification problems.

2.5.2. Macro-F1 Score : The Macro F1 score is calculated to tackle the problem of multi-class accuracy. This score allows equal importance to classes while evaluating. F1-score for class k is defined as follows:

$$F1_k = \frac{2 \cdot \text{Precision}_k \cdot \text{Recall}_k}{\text{Precision}_k + \text{Recall}_k} \tag{19}$$

The Macro-F1 score is then calculated as the unweighted average over all K classes.:

$$\text{Macro} - \text{F1} = \frac{1}{K} \sum_{k=1}^{K} F1_k \tag{20}$$

In case of multi-class classification problems, this measure is ideal since it treats all classes equally.

2.6. Temporal Stability

Temporal stability analysis assesses the model's performance over time. In this study, the absolute accuracy difference between the two-time points measures stability (2018 and 2020):

$$\text{Stability} = \text{Accuracy}_{2018} - \text{Accuracy}_{2020} \tag{21}$$

If this measure has a smaller value, it implies greater temporal robustness and generalisation corresponding to the model.

2.7. Variable Importance

Permutation-based importance assessments are employed to determine the importance of predictors in relation to model performance. What Variable jis teaches us about significance:

$$VI_j = \text{Performance}_{\text{original}} - \text{Performance}_{\text{permuted}} \tag{22}$$

The original data set is used to evaluate the model performanceoriginal. Performancepermuted refers to the performance that arises from the random permutation of values of variable j. More reduction in performance is proportional to more important variable.

3. Methodological Framework

3.1. Data Description

The data employed in this study is sourced from the official annual statistical reports published by the Ministry of Health and Environment, Republic of Iraq. Data for 2018 and 2020 were specifically retrieved from the website of the Directorate of Planning.: https://planning.moh.gov.iq/. The reports give comprehensive and trustworthy health statistics with respect to the governorates, cancer cases, demographic features and distribution of the population. The dataset was assembled and organized for the purpose of performing spatial analysis and classification modelling. The following table describes the included variables of the study, namely the demographic and health-related variables to analyse the spatial distribution of childhood cancer in Iraqi governorates.

Table 1. Descriptive Statistics of Study Variables

Variable	Description	Variable	Description
x1	Year	x9	Total female population below 15 years
x2	Governorate	x10	Total male population below 15 years
x3	Children with cancer aged 0–4 years	x11	Annual cancer incidence rate among females per 100,000 female population
x4	Children with cancer aged 5–9 years	x12	Annual cancer incidence rate among male per 100,000 female population
x5	Children with cancer aged 10–15 years	x13	Annual cancer incidence rate among Total per 100,000 female population
x6	Total number of childhood cancer cases	x14	Population growth rate
x7	Percentage of Total Population	x15	Dependency ratio
x8	Total population below 15 years	x16	Median age

3.2. Spatial Epidemiology

Table (1) show the comparison of variable importance between 2018 and 2020 highlights dynamic shifts in the key determinants of disease incidence across different dependent variables and time periods, while also revealing a set of consistently influential predictors. Notably, variable x6 maintained a dominant position across most models, particularly for x4 and x5, indicating its robustness as a core explanatory factor across age groups. Within the framework of spatial epidemiology, these findings emphasize the role of spatial heterogeneity and temporal variation in shaping disease patterns. The observed transition in the importance of predictors for x3, where influence shifted from x6 in 2018 to x7 and x8 in 2020, reflects changes in the underlying spatial structure of the data. Furthermore, the persistent significance of spatially derived features such as spatial lag (x6_lag) and local differences (x8_diff, x7_diff) confirms the importance of incorporating spatial spillover effects and localized variations. From a methodological perspective, the integration of machine learning techniques with spatial feature engineering provides a flexible framework capable of capturing complex, nonlinear relationships and spatial

dependencies that are often overlooked in traditional statistical models. The increasing importance of difference-based spatial features in 2020, particularly for x5, suggests that local spatial variability plays a growing role in explaining disease incidence among older age groups. Overall, this integrated approach enhances the analytical capacity of spatial epidemiology by combining data-driven modeling with explicitly constructed spatial features, thereby improving both predictive performance and interpretability. These results demonstrate that leveraging machine learning within a spatially informed framework can effectively uncover evolving patterns of disease distribution and provide deeper insights into spatial health dynamics.

Table 2. Comparison of the Top 10 Most Important Variables Between 2018 and 2020

Rank	x3 (2018)	Importance	x3 (2020)	Importance	x4 (2018)	Importance	x4 (2020)	Importance	x5 (2018)	Importance	x5 (2020)	Importance
1	x6	0.06	x7	0.037	x6	0.0316	x6	0.0386	x6	0.0297	x6	0.043
2	x6_diff	0.029	x8	0.036	x15	0.0208	x8	0.0157	x12	0.0178	x8_diff	0.0126
3	x13_lag	0.011	x6	0.017	x11_lag	0.0187	x7	0.0072	x12_diff	0.0057	x7_diff	0.0111
4	x11_lag	0.008	x6_lag	0.015	x6_diff	0.0165	x15	0.0055	x6_diff	0.0056	x7	0.0107
5	x9	0.006	x8_diff	0.008	x13_lag	0.0069	x15_diff	0.0055	x13_diff	0.0037	x6_diff	0.0083
6	x10	0.005	x8_lag	0.007	x14_lag	0.0057	x10	0.0042	x11_diff	0.0037	x8	0.006
7	x8	0.005	x10_lag	0.006	x8	0.0022	x6_lag	0.0027	x10	0.0032	x16_lag	0.0048
8	x16_lag	0.004	x7_lag	0.005	x9_diff	0.002	x12_lag	0.0027	x9	0.0029	x9_lag	0.0047
9	x10_diff	0.004	x7_diff	0.004	x9	0.0012	x16_lag	0.0027	x9_diff	0.0028	x8_lag	0.0036
10	x8_diff	0.003	x16_diff	0.003	x7_diff	0.001	x8_lag	0.0023	x16_diff	0.0027	x6_lag	0.0032

3.3. Evaluation of Best-Performing Machine Learning Models Across Time

Table (3) outlines the best-performing models and their corresponding accuracy rates (with 95% confidence intervals) for variables x3, x4, and x5 in 2018 and 2020. XGBoost consistently outperformed other models, securing the top position in most evaluations. The highest individual accuracy was recorded for variable x3 in 2018 using Random Forest (72.2%), though its performance dropped significantly in 2020 (44.4%). On the other hand, variables x4 and x5 showed improved accuracy over time, rising from 50.0% to 66.7% for x4, and from 55.6% to 61.1% for x5. In 2020, Random Forest tied with XGBoost as the top model for x5.

Table 3. Best Model Performance and Accuracy Shifts Across 2018 and 2020

Variable	Best Model (2018)	Accuracy (95% CI)	Best Model (2020)	Accuracy (95% CI)
x3	Random Forest	72.2% (46.5–90.3)	XGBoost	44.4% (21.5–69.2)
x4	XGBoost	50.0% (26.0–74.0)	XGBoost	66.7% (41.0–86.7)
x5	XGBoost	55.6% (30.8–78.5)	Random Forest / XGBoost	61.1% (35.8–82.7)

3.4. Temporal Stability Comparison of Classification Models Between 2018 and 2020

The table(4) & Figure (1) presents the results of comparing the stability of classification models between 2018 and 2020 using a stability index based on the absolute difference in accuracy ($|\Delta\text{Accuracy}|$). This comparison is conducted for three dependent variables representing age-specific categories of disease incidence: x3 (children under 5 years), x4 (children aged 5 to under 10 years), and x5 (children aged 10 to under 15 years). A lower value of the stability index indicates more consistent model performance over time.

The results reveal that model stability varies depending on the dependent variable. For x3, the models exhibited noticeable variability in stability. Both Random Forest and SVM recorded the highest instability values (**??**), indicating substantial fluctuations in performance between 2018 and 2020. In contrast, Shrinkage-QDA and XGBoost demonstrated relatively better stability, each with a value of 0.222, suggesting more consistent performance across the two years compared to the other models.

For x4, the models showed a generally higher level of stability compared to x3. The Random Forest model achieved the best stability with a value of 0.111, while the remaining models (Shrinkage-QDA, SVM, and XGBoost) exhibited similar values of approximately 0.167, indicating a moderate level of temporal consistency in performance.

Regarding x5, the XGBoost model emerged as the most stable among all models, recording the lowest stability index value 0.056, which reflects highly consistent performance between 2018 and 2020. In contrast, SVM showed the highest instability 0.278, indicating greater variability in performance. Meanwhile, Random Forest and Shrinkage-QDA demonstrated moderate stability levels, with values of 0.167 and 0.222, respectively.

These findings suggest that model stability is not uniform across all age groups, but rather depends on the nature of the dependent variable and the underlying statistical and spatial structure of the data. Notably, tree-based models—particularly XGBoost—demonstrated a stronger ability to maintain stable performance over time, especially for x5, whereas other models, such as SVM, exhibited greater variability in certain cases.

It is also important to note that the relatively small sample size, represented by the limited number of spatial units, may contribute to increased variability in model performance across time periods, particularly when a large number of original and spatially derived variables are included. Nevertheless, the results indicate that advanced models are capable of achieving a better balance between accuracy and stability, thereby enhancing their reliability in analyzing spatial patterns of disease incidence.Based on the stability index, it can be concluded that XGBoost is the most suitable model in terms of temporal stability, particularly for x5, while Random Forest demonstrates strong stability performance for x4, with noticeable variation in the stability of other models across different dependent variables.

Table 4. Temporal Stability of Classification Models (2018 vs. 2020)

Dependent Variable	Model	Accuracy (2018)	Macro-F1 (2018)	Accuracy (2020)	Macro-F1 (2020)	Stability Index
x3	Random Forest	0.722	0.868	0.389	0.571	0.333
x3	Shrinkage-QDA	0.444	0.523	0.222	0.615	0.222
x3	SVM	0.389	0.386	0.056	0.2	0.333
x3	XGBoost	0.667	0.823	0.444	0.617	0.222
x4	Random Forest	0.444	0.437	0.333	0.342	0.111
x4	Shrinkage-QDA	0.333	0.337	0.5	0.481	0.167
x4	SVM	0	—	0.167	0.286	0.167
x4	XGBoost	0.5	0.494	0.667	0.669	0.167
x5	Random Forest	0.444	0.438	0.611	0.636	0.167
x5	Shrinkage-QDA	0.278	0.274	0.5	0.53	0.222
x5	SVM	0.167	0.248	0.444	0.615	0.278
x5	XGBoost	0.556	0.556	0.611	0.654	0.056

3.5. Model Comparison Based on Misclassification

As the table indicates, misclassification rates and accuracy of models for the three variable (x3, x4 and x5) of 2018 and 2020 differ considerably. This is due to performances and stability of models. In 2018, for the most part, Random Forest and XGBoost model achieved the highest accuracy and lowest misclassification rates. The Random Forest method got the highest accuracy performance of (0.7222) and misclassification (0.2778) in x3, followed by XGBoost. Shrinkage-QDA and SVM gave even larger error rates, the results showed. Similarly, for x5, the highest accuracy rate was achieved by XGBoost with SVM having the highest misclassification rates, showing it was unsuitable. The dominance of tree-based models was confirmed in 2020 as XGBoost performed best on x4 (Accuracy=0.6667) while Random Forest and XGBoost performed best on x5 (Accuracy=0.6111, Misclassification=0.3889). Conversely, among all variables in both years, SVM has the highest misclassification

Figure 1. Temporal Stability of Classification Models Across Variables (x3, x4, x5) in Iraq (2018–2020).

rates by 0.9444 (x3 2020) and 1.0 (x4 2018) indicates that it cannot grasp the underlying structure of the data. The performance of shrinkage-qda was moderate and there was a little improvement in 2020 for the x4x5. Yet, it did not surpass the performance of any tree-based model. As reflected in results, Random Forest and XGBoost models are more robust and reliable, with lower classification errors, especially suited for complex and non-linear classification problems. Among the models examined, the SVM was the worst performer while Shrinkage-QDA gave somewhere in between. The findings indicate that appropriate model selection should be made to help reduce misclassification with respect to spatial and heterogeneous data.

Table 5. Comparison of Classification Models Based on Misclassification Rates and Accuracy for Variables (x3, x4, x5) in Iraq (2018–2020)

Year	Dependent	Model	Total	Correct	Misclassified	Accuracy	Misclassification Rate
2018	x3	Shrinkage_QDA	18	8	10	0.4444	0.5556
2018	x3	SVM	18	7	11	0.3889	0.6111
2018	x3	Random Forest	18	13	5	0.7222	0.2778
2018	x3	XGBoost	18	12	6	0.6667	0.3333
2018	x4	Shrinkage_QDA	18	6	12	0.3333	0.6667
2018	x4	SVM	18	0	18	0	1
2018	x4	RandomForest	18	8	10	0.4444	0.5556
2018	x4	XGBoost	18	9	9	0.5	0.5
2018	x5	Shrinkage_QDA	18	5	13	0.2778	0.7222
2018	x5	SVM	18	3	15	0.1667	0.8333
2018	x5	RandomForest	18	8	10	0.4444	0.5556
2018	x5	XGBoost	18	10	8	0.5556	0.4444
2020	x3	Shrinkage_QDA	18	4	14	0.2222	0.7778
2020	x3	SVM	18	1	17	0.0556	0.9444
2020	x3	RandomForest	18	7	11	0.3889	0.6111
2020	x3	XGBoost	18	8	10	0.4444	0.5556
2020	x4	Shrinkage_QDA	18	9	9	0.5	0.5
2020	x4	SVM	18	3	15	0.1667	0.8333
2020	x4	RandomForest	18	6	12	0.3333	0.6667
2020	x4	XGBoost	18	12	6	0.6667	0.3333
2020	x5	Shrinkage_QDA	18	9	9	0.5	0.5
2020	x5	SVM	18	8	10	0.4444	0.5556
2020	x5	RandomForest	18	11	7	0.6111	0.3889
2020	x5	XGBoost	18	11	7	0.6111	0.3889

3.6. Spatial Comparison

The figures (2,3,4,5,6,7) illustrates the spatial distribution of classification results across Iraqi governorates using four machine learning models (Shrinkage-QDA, SVM, Random Forest, and XGBoost) for the three dependent variables (x3, x4, x5) in the years 2018 and 2020.A clear variation in spatial patterns can be observed across the models. Specifically, the Shrinkage-QDA and SVM models generate relatively smooth and coherent spatial distributions, where neighboring governorates tend to belong to similar classes, reflecting stable and gradual spatial transitions. In contrast, the Random Forest and XGBoost models produce more fragmented and heterogeneous spatial patterns, characterized by scattered high and low classifications, indicating their stronger ability to capture nonlinear relationships and complex spatial interactions.With respect to the variables, the maps of x3 exhibit relatively consistent and balanced spatial patterns across all models, with limited discrepancies in classification. In comparison, x4 demonstrates greater spatial variability, particularly in the northern and southern regions, where noticeable differences between models emerge. The variable x5 shows the highest level of spatial complexity, with substantial inconsistencies in classification results across models, especially in western and southern governorates, indicating higher sensitivity to model structure.From a temporal perspective, the comparison between 2018 and 2020 reveals a general degree of spatial stability in the results of Shrinkage-QDA and SVM models. However, the tree-based models (Random Forest and XGBoost) exhibit more pronounced temporal changes, suggesting that these models are more responsive to variations in the underlying data over time.Overall, the figure demonstrates that the choice of model significantly affects not only classification accuracy but also the spatial representation of results, highlighting the importance of incorporating spatial interpretation alongside statistical evaluation in health-related studies.

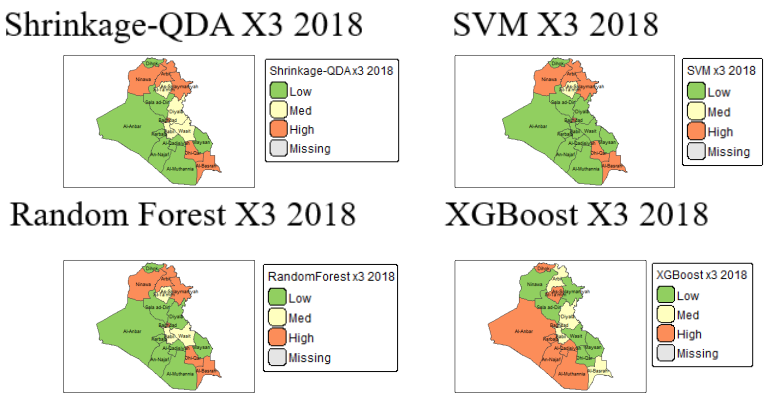

Figure 2. Spatial Comparison of Classification Performance Across Machine Learning Models (Shrinkage-QDA, SVM, Random Forest, and XGBoost) for Variable (X3) in Iraq, 2018.

3.7. Research Gap

Despite the growing body of literature in spatial epidemiology and machine learning applications, several critical limitations remain. Most existing studies have focused on either traditional spatial statistical models or standalone machine learning approaches, without fully integrating both frameworks. In particular, previous research has typically examined spatial effects through conventional methods such as spatial regression or disease mapping, while machine learning studies have often ignored explicit spatial dependencies or treated spatial information implicitly.

Moreover, the majority of studies have addressed only a single analytical dimension, such as prediction accuracy or spatial pattern detection, without simultaneously considering classification performance, spatial feature engineering, and temporal consistency. Very limited attention has been given to the incorporation of spatially derived features—such as spatial lag and local difference variables—within classification frameworks.

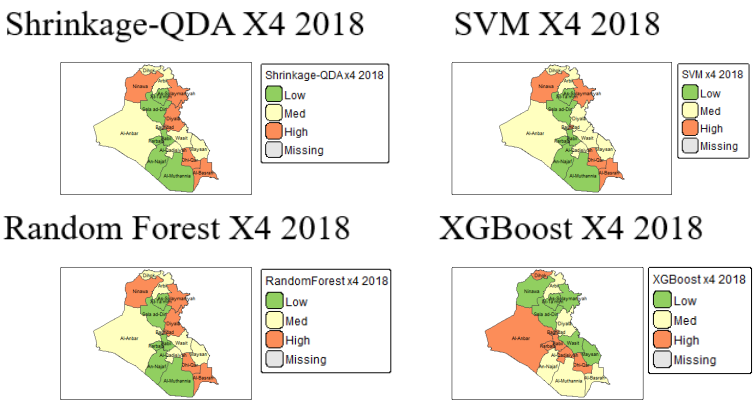

Figure 3. Spatial Comparison of Classification Performance Across Machine Learning Models (Shrinkage-QDA, SVM, Random Forest, and XGBoost) for Variable (X4) in Iraq, 2018.

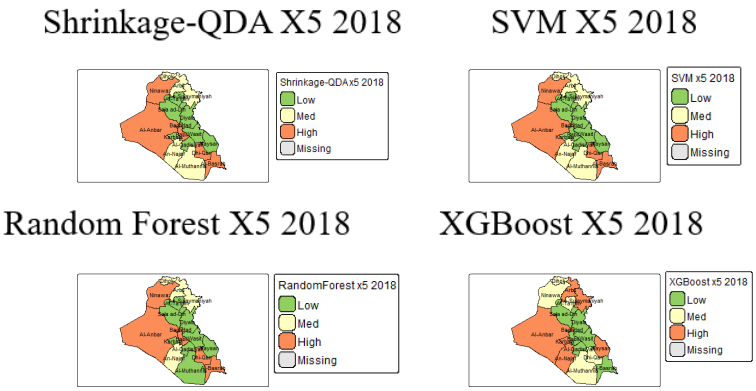

Figure 4. Spatial Comparison of Classification Performance Across Machine Learning Models (Shrinkage-QDA, SVM, Random Forest, and XGBoost) for Variable (X5) in Iraq, 2018.

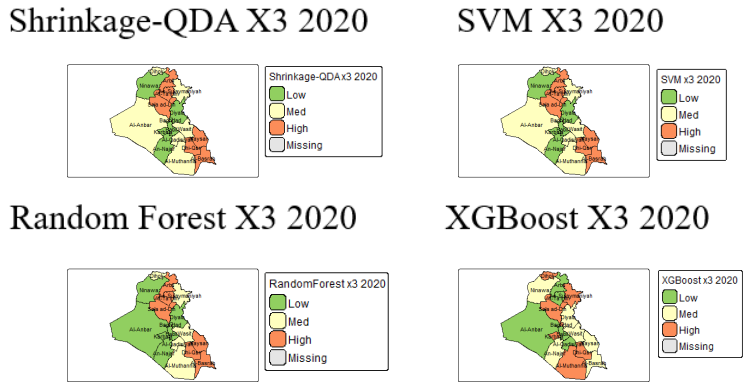

Figure 5. Spatial Comparison of Classification Performance Across Machine Learning Models (Shrinkage-QDA, SVM, Random Forest, and XGBoost) for Variable (X3) in Iraq, 2020.

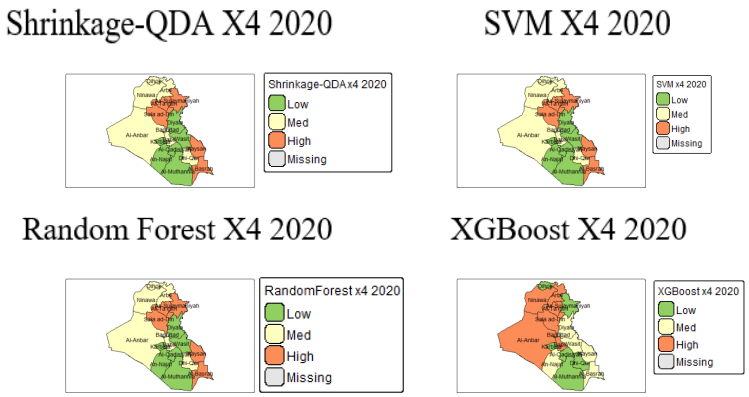

Figure 6. Spatial Comparison of Classification Performance Across Machine Learning Models (Shrinkage-QDA, SVM, Random Forest, and XGBoost) for Variable (X4) in Iraq, 2020.

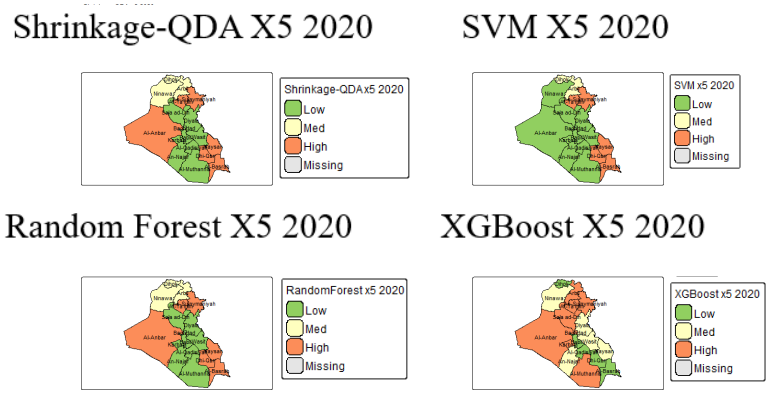

Figure 7. Spatial Comparison of Classification Performance Across Machine Learning Models (Shrinkage-QDA, SVM, Random Forest, and XGBoost) for Variable (X5) in Iraq, 2020.

Importantly, temporal stability analysis remains largely overlooked in the literature. Most studies evaluate model performance at a single time point, without assessing the robustness of models across different time periods. This limitation reduces the reliability of findings, particularly in dynamic spatial datasets where patterns may evolve over time.

Therefore, there is a clear research gap in developing an integrated framework that combines machine learning, spatial feature engineering, classification modeling, and temporal stability analysis to provide a more comprehensive and reliable understanding of spatiotemporal disease patterns.

3.8. Contribution of the Study

Research done on the theme under investigation where not enough research has been conducted in South Africa. We propose a new hybrid model for analyzing disease incidences that uses spatio-phylogenic features along with machine learning.

The major contributions of this study may be summarized as follows.

Scientists will create a framework that combines data-driven machine learning prediction with human expert spatially explict feature construction.

The process of spatial feature engineering will include:

Reflecting neighborhood effects, spatial lag variables.
Devices to capture local differences local heterogeneity.

Execution of classification modeling of disease incidence for three age groups (x3,x4,x5) which helps in studying the age-wise spatial patterns better.

A new stability index to assess temporal stability has been introduced to measure the reliability of performance over different time slices (2018-2020) Based on the absolute difference in accuracy ($|\Delta\text{Accuracy}|$), the stability index is.

Comprehensive comparison between models (Random Forest, SVM, XGBoost, Shrinkage-QDA) to find the most efficient and stable model for identifying and describing spatiotemporal variability.

Spatial statistical theory and machine learning data-driven approaches for improved spatial epidemiology analysis.
In short, this study offers a novel and comprehensive methodological contribution by putting spatial and/or temporal and machine learning dimensions into one classification framework that will yield more robust insights into disease dynamics which can be interpreted more easily.

4. Conclusion

As per results, the dependent variables were efficiently converted into three classes low, medium, and high using tertiles to simplify data and retain the spatial variability. After the transformation, it boosted the capacity to employ classification models and compare ageing processes. Also, a better understanding of spatial variation was obtained through the application of spatial lag and spatial difference variables to capture the neighboring effect and the local deviation effect. The models are then considered to be more realistic than other existing choice models.
The investigation also reveals some constant factors and some factors that are variable. Always an important variable, x6's contribution has consistently remained unchanged over different periods of time. Its contrary to other variables, especially variable x3 which varies in importance through time.
Taking stability through time into account, the XGBoost model came out as the winner as it was the most stable model we analyzed, especially for x5, while the SVM models are much more volatile throughtime Additionally, when the outputs of the model were assessed spatially, the tree-based models in particular Random Forest and XGBoost captured spatial heterogeneity much better.
The space comparison showed that the classification models behaved differently in the vicinity. The Random Forest model's success in achieving spatial equilibrium in class distribution affirms its adequate representation of geographical gradients in the data. On the other hand, XGBoost strongly overestimated the class of high risk in certain areas due to spatial overestimation. Over time, the SVM model is likely to force the high-risk category into a few provinces, indicating sensitivity to outliers. In contrast, the Shrinkage-QDA model produced smoother spatial patterns - somewhat more conservative and more limited in expressing these spatial variations.
The findings suggest that machine learning integrated with spatial feature engineering provides a robust framework for analysing spatial health data. Deciding which model to use not only affects how accurately certain aspects of my prediction might become, but also its geography. Even when Random Forest produces reliable and interpretable results, XGBoost is more powerful and temporally stable. However, Random Forest is more spatially balanced.

REFERENCES

1. M. Amanulloh, et al., *Comparative analysis of machine learning models for spatial classification*, International Statistical Journal, vol. 15, no. 2, pp. 120–135, 2026.
2. L. Anselin, *Spatial Econometrics: Methods and Models*, Kluwer Academic Publishers, 1988.
3. L. Breiman, *Random forests*, Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
4. T. Chen, and C. Guestrin, *XGBoost: A scalable tree boosting system*, Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), 2016.
5. R. Chowdhury, et al., *Spatio-temporal machine learning approaches for disease modeling*, PLoS Neglected Tropical Diseases, vol. 19, no. 3, e0012800, 2025.
6. C. Cortes, and V. Vapnik, *Support-vector networks*, Machine Learning, vol. 20, pp. 273–297, 1995.

7. J. Friedman, *Regularized discriminant analysis*, Journal of the American Statistical Association, vol. 84, no. 405, pp. 165–175, 1989.
8. T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed., Springer, 2009.
9. G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*, 2nd ed., Springer, 2021.
10. K. Kianfar, et al., *The future of spatial epidemiology in the AI era: Enhancing machine learning approaches with explicit spatial structure*, Geospatial Health, vol. 20, no. 1, p. 1386, 2025.
11. M. Kuhn, and K. Johnson, *Applied Predictive Modeling*, Springer, 2013.
12. J. LeSage, and R. K. Pace, *Introduction to Spatial Econometrics*, CRC Press, 2009.
13. Y. Liu, et al., *Integrating machine learning and spatial analysis for public health prediction*, BMC Public Health, vol. 25, p. 22077, 2025.
14. M. Noaeen, et al., *Machine learning for spatial health prediction: A comparative study*, Scientific Reports, vol. 16, p. 34287, 2026.
15. S. Sahat, et al., *Machine learning-based spatial modeling of disease incidence using Random Forest and XGBoost*, Environmental Health Research, vol. 18, no. 4, p. 23119, 2025.
16. M. Sokolova, and G. Lapalme, *A systematic analysis of performance measures for classification tasks*, Information Processing & Management, vol. 45, no. 4, pp. 427–437, 2009.
17. D. S. Wilks, *Statistical Methods in the Atmospheric Sciences*, Academic Press, 2011.
18. X. Zhang, et al., *Enhancing spatial prediction using spatial lag features in machine learning models*, ISPRS International Journal of Geo-Information, vol. 13, no. 9, p. 330, 2024.