Supplementary

Table S1 Parameters for the data balancing algorithm

Method	Parameter
Adaptive Synthetic (ADASYN)	baseClass = no, dist = HEOM, beta =0.15
Synthetic Minority Over-Sampling Technique for Regression with Gaussian Noise (SMOGN)	dist = HEOM, C.perc = list (negative =1.5, positive =2.0)
Synthetic Minority Over-Sampling Technique (SMOTE)	dist = HEOM, C.perc = list (negative =1.2, positive =3.0)

Table S2 Summary of machine learning algorithms

Model	Summary
Elastic net regression	Elastic net regression is a widely employed linear regression technique that combines the characteristics of Ridge regression and Lasso regression. By merging the regularization terms of these two methods, elastic net regression aims to overcome the limitations of using Ridge regression or Lasso regression individually when dealing with data sets exhibiting multicollinearity, high feature dimensionality relative to sample size, or strong correlations among certain features
Random forest	Random forest is an extension of the Bagging ensemble learning method, widely employed in both classification and regression tasks. Its base learner is fixed as a decision tree, constructing multiple decision trees using random subsets of data and variables. Each decision tree serves as a classifier, collectively forming a forest, and the random forest integrates all classification votes, designating the most frequently voted class as the final output result
Support vector machine	Support Vector Machine is a commonly employed supervised learning algorithm for classification and regression analysis. By considering a sample set containing positive and negative instances, SVM aims to identify a hyperplane that effectively separates data points of different classes, maximizing the distance between data points of each class and the hyperplane. This process enhances the model's generalization capability
XGBoost	XGBoost is an efficient gradient boosting decision tree algorithm based on the idea of Boosting. It integrates multiple weak learners into a strong learner through a specific method, making decisions jointly through multiple decision trees. The prediction of each tree is the difference between the target value and the sum of all previous tree predictions. By aggregating all results, the final outcome is obtained, thereby enhancing the overall effectiveness of the model
Artificial Neural Network	Neural networks can amalgamate the predictive outcomes of multiple independent neural networks by integrating them through averaging or weighted averaging. This ensemble learning technique aids in reducing overfitting, thereby enhancing the robustness and generalization capabilities of the model, especially beneficial for handling complex datasets or high-dimensional feature spaces. Neural network models accomplish feature extraction and learning by learning these connection weights and are applied to tasks such as classification or regression

https://dx.doi.org/10.21037/jgo-24-777

Table S3 Hyperparameters for the final models

Model	Hyperparameter
ADASYN	
Elastic net regression	alpha =0.1, lambda =0.001668101
Random forest	mtry =5, splitrule =gini, min.node.size =5
Support vector machine	degree =2, scale =1, C =1
XGBoost	nrounds =100, max_depth =3, eta =0.1, gamma =0, colsample_bytree =0.5, min_child_weight =2, subsample =0.7
Artificial Neural Network	Size =20, decay =0.2, bag =0.8
ADASYN (one hot encoding)	
Elastic net regression	alpha =0.5, lambda =0.02154435
Random forest	mtry =5, splitrule =factor, min.node.size =5
Support vector machine	degree =1, scale =10, C =0.01
XGBoost	nrounds =200, max_depth =3, eta =0.05, gamma =0.4, colsample_bytree =0.5, min_child_weight =1, subsample =0.8
Neural Network	Size =15, decay =0.2, bag =0.6
SMOGN	
Elastic net regression	alpha =0.5, lambda =0.001668101
Random forest	mtry =4, splitrule =gini & extratrees, min.node.size =5
Support vector machine	degree =2, scale =10, C =1
XGBoost	nrounds =200, max_depth =3, eta =0.1, gamma =0, colsample_bytree =0.5, min_child_weight =1, subsample =0.7
Artificial Neural Network	Size =10, decay =0.1, bag =0.6
SMOGN (one hot encoding)	
Elastic net regression	alpha =0.3, lambda =0.001668101
Random forest	mtry =4, splitrule =gini & factor, min.node.size =5
Support vector machine	degree =2, scale =1, C =0.1
XGBoost	nrounds =200, max_depth =3, eta =0.1, gamma =0.3, colsample_bytree =0.7, min_child_weight =1, subsample =0.9
Artificial Neural Network	Size =10, decay =0.2, bag =0.8
SMOTE	
Elastic net regression	alpha =0.2, lambda =0.001668101
Random forest	mtry =5, splitrule =gini & extratrees, min.node.size =5
Support vector machine	degree =2, scale =1, C =1
XGBoost	nrounds =200, max_depth =3, eta =0.1, gamma =0.2, colsample_bytree =0.7, min_child_weight =1, subsample =0.7
Artificial Neural Network	Size =10, decay =0.1, bag =0.7
SMOTE (one hot encoding)	
Elastic net regression	alpha =0.1, lambda =0.02154435
Random forest	mtry =5, splitrule =gini & extratrees, min.node.size =5
Support vector machine	degree =2, scale =1, C =0.1
XGBoost	nrounds =200, max_depth =3, eta =0.1, gamma =0.3, colsample_bytree =0.7, min_child_weight =1, subsample =0.8
Artificial Neural Network	Size =15, decay =0.1, bag =0.8

https://dx.doi.org/10.21037/jgo-24-777

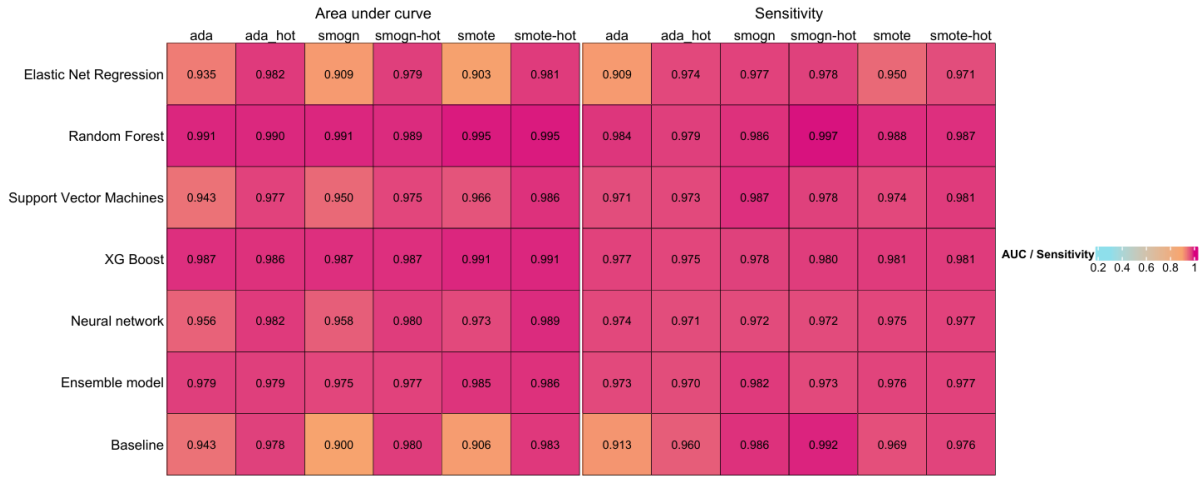

Figure S1 Heatmap of model performance for lymph node metastasis prediction of GISTs for each machine learning algorithm in Training Group. GIST, gastrointestinal stromal tumors; ROC, receiver operator characteristic; AUC, area under the ROC curve; ada, Adaptive Synthetic Sampling; SMOTE, synthetic minority over-sampling technique; SMOGN, synthetic minority over-sampling technique for regression with Gaussian noise; ada-hot, Adaptive Synthetic Sampling and one-hot encoding; SMOTE-hot, synthetic minority over-sampling technique and one-hot encoding; SMOGN-hot, synthetic minority over-sampling technique for regression with Gaussian noise and one-hot encoding.

https://dx.doi.org/10.21037/jgo-24-777

Table S4 Clinicopathological characteristics of patients with GIST after PSM

Characteristics	No lymph node metastasis (n=172)	Lymph node metastasis (n=42)	P value
Age (years)	61.30±12.80	58.67±18.05	0.275
Race			0.736
White	125 (72.7)	33 (78.6)	
Black	32 (18.6)	6 (14.3)	
Others	15 (8.7)	3 (7.1)	
Primary site			0.838
Stomach	96 (55.8)	24 (57.1)	
Small intestine	56 (32.6)	15 (35.7)	
Colon	9 (5.2)	1 (2.4)	
Other	11 (6.4)	2 (4.8)	
Tumor size (cm)			0.952
≥2 and <5	26 (15.1)	5 (11.9)	
≥5 and ≤10	59 (34.3)	15 (35.7)	
>10	84 (48.8)	21 (50.0)	
Unknown	3 (1.7)	1 (2.4)	
UICC TNM stage			0.754
Stage 1	31 (18.0)	10 (23.8)	
Stage 2	20 (11.6)	6 (14.3)	
Stage 3	25 (14.5)	5 (11.9)	
Stage 4	23 (13.4)	7 (16.7)	
Unknown	73 (42.4)	14 (33.3)	
Mitotic index			0.275
<5/50	68 (39.5)	15 (35.7)	
5/50–10/50	20 (11.6)	6 (14.3)	
>10/50	33 (19.2)	13 (31.0)	
Unknown	51 (29.7)	8 (19.0)	
Summary stage			0.15
Distant	124 (72.1)	35 (83.3)	
Localized	21 (12.2)	1 (2.4)	
Regional	27 (15.7)	6 (14.3)	
Total number of regional lymph nodes	8.59 ±10.94	9.74±10.28	0.537
NIH risk category			0.793
Low	13 (7.6)	2 (4.8)	
Intermediate	6 (3.5)	1 (2.4)	
High	97 (56.4)	27 (64.3)	
Unknown	56 (32.6)	12 (28.6)	
M stage			0.079
I	56 (32.6)	21 (50.0)	
0	114 (66.3)	20 (47.6)	
Unknown	2 (1.2)	1 (2.4)	

Data are presented as n (%) or mean ± SD. AJCC, American Joint Committee on Cancer; GIST, gastrointestinal stromal tumor; NIH, National Institutes of Health; PSM, propensity score matching; SD, standard deviation; TNM, Tumor Node Metastasis; UICC, Union for International Cancer Control.

https://dx.doi.org/10.21037/jgo-24-777

Table S5 Importance of variable

Variables	Contribution to the model
M stage I	100
Summary stage (localized)	76.33
Age	72.04
Primary site stomach	67.75
Summary stage (distant)	61.92
Tumor size >10 cm	58.29
UICC TNM stage IV	52.36
Summary stage (regional)	50.98
NIH risk category (high)	47.81
Primary site small intestine	42.19
Tumor size ≥5 and ≤10 cm	39.11
UICC TNM stage II	35.70
Mitotic index >10/50	33.44
UICC TNM stage III	31.55
Tumor size ≥2 and <5 cm	28.17
Mitotic index <5/50	25.68
NIH risk category (intermediate)	22.49
Mitotic index 5/50–10/50	19.62
M stage 0	18.73
No neoadjuvant therapy	15.27

NIH, National Institutes of Health; TNM, Tumor Node Metastasis; UICC, Union for International Cancer Control.

https://dx.doi.org/10.21037/jgo-24-777