ORIGINAL ARTICLE

European Journal of BREAST HEALTH

DOI: 10.4274/ejbh.galenos.2026.2026-1-10
Eur J Breast Health 2026;22(3):307-317

Development and Internal Validation of a Clinical Data-Based Machine Learning Web Calculator for Predicting Recurrence in Granulomatous Lobular Mastitis: A Multicenter Retrospective Study

Jiao Feng[1], Ruiyang Wu[2], Jin Chen[2], Yi Li[1]

[1]Department of Gastrointestinal Surgery, The Affiliated Chengdu 363 Hospital of Southwest Medical University, Chengdu, China
[2]Department of Breast and Thyroid, Sichuan Provincial Hospital for Women and Children (Affiliated Women and Children's Hospital of Chengdu Medical College), Chengdu, China

ABSTRACT

Objective: Granulomatous lobular mastitis (GLM) is a disease characterized by a high recurrence rate and the absence of a standard treatment, making prognostic prediction crucial. While promising, existing machine learning models are limited by single-center data and small sample sizes. This study aimed to develop and validate machine learning models using a large multicenter dataset to predict GLM recurrence and build a clinical web calculator.

Materials and Methods: In this retrospective cohort study, data from 318 GLM patients at two tertiary hospitals (diagnosed between 2019 and 2024) were used to train and evaluate five machine learning models. Performance was assessed by accuracy, area under the curve (AUC), F1-score, sensitivity, and specificity.

Results: The five models demonstrated comparable discriminatory performance, with AUCs ranging from 0.778 to 0.808 and no statistically significant differences among them. Among them, random forest (RF) excelled in composite and sensitivity metrics (F1 score: 0.639; accuracy: 76.2%; sensitivity: 50%), whereas logistic regression achieved the top AUC (0.808), and the support vector machine achieved the best specificity (95.3%). Based on its balanced performance across multiple metrics, RF was selected for deployment to develop a publicly accessible web application platform (https://w12251393.shinyapps.io/predictGLM/). In the RF model, white blood cell count emerged as the top predictor, followed by age at diagnosis, the origin of the primary tumor, surgical excision, antitubercular therapy, corticosteroid therapy, and abscess drainage, in descending order of importance.

Conclusion: Although retrospective in design, this study developed a multicenter RF model and implemented it as an accessible web calculator, providing a valuable tool for personalized recurrence prediction and treatment decision-making in GLM. The model should be used as a risk stratification aid to support clinical decision-making rather than as a definitive predictive instrument.

Keywords: Granulomatous mastitis; machine learning; recurrence

Corresponding Author: Yi Li MD;
E-mail: wcfj123123@163.com **ORCID:** orcid.org/0009-0005-5262-1030

Received: 31.01.2026 **Accepted:** 15.03.2026 **Available Online Date:** 17.06.2026

Cite this article as: Feng J, Wu R, Chen J, Li Y. Development and internal validation of a clinical data-based machine learning web calculator for predicting recurrence in granulomatous lobular mastitis: a multicenter retrospective study. Eur J Breast Health. 2026;22(3):307-317

KEY POINTS

- This study addresses the high recurrence rate of granulomatous lobular mastitis by developing the first machine-learning-based online calculator (random forest) for predicting recurrence.
- The model, constructed from multicenter data, demonstrates balanced predictive performance and identifies seven key predictors, including white blood cell count and surgical intervention.
- The result has been translated into a freely accessible web tool that provides instant, individualized recurrence-risk assessments to support clinical decision-making and to serve as a risk-stratification aid.

Introduction

Granulomatous lobular mastitis (GLM) was a non-puerperal chronic inflammatory breast disease characterized by non-caseating granulomas and microabscesses confined to the breast lobules (1-3). Based on a significant rise in reported cases over the past decade, the incidence of GLM has increased dramatically (4). GLM typically presented with painful breast masses, redness, and abscesses that could progress to fistulas and sinus tracts, often leading to significant breast deformity and a high recurrence rate (5-7).

Given its notoriously high recurrence rate, reported in various studies to range from 24% to 40% (8-11), GLM has often been referred to as "incurable cancer" (12). This combination of a high relapse risk and the lack of standardized therapeutic guidelines makes accurate prognostic assessment crucial (13). Although some studies attempted to employ staging systems to predict patient prognosis (12, 14, 15), their predictive efficacy was largely unsatisfactory due to the disease's marked heterogeneity. For instance, in our previous work, we developed the 1st edition of GLM stage for predicting GLM prognosis; however, it yielded a suboptimal area under the curve (AUC) of only 0.642 (12). This limitation of conventional approaches highlights the need for more advanced methods, such as machine learning, which may better capture complex patterns in heterogeneous diseases. Consequently, recent research has increasingly leveraged machine learning models to integrate complex, multi-dimensional patient data, demonstrating promising predictive performance for recurrence risk (16, 17). For instance, Li et al. (16) demonstrated that the neural network model achieved higher predictive accuracy. These promising findings, however, faced substantial clinical implementation barriers. The models had not been integrated into clinical workflows for direct assessment, and their development was constrained by small, single-center datasets with incomplete metrics, which collectively limited their generalizability and immediate practical utility. To overcome these limitations, this study leveraged a large-scale, multi-center dataset to develop and compare multiple machine learning models, with the specific objective of building a web-based clinical calculator for prognostic assessment.

Materials and Methods

Data and Samples

This retrospective cohort study initially enrolled 599 patients diagnosed with GLM from two tertiary care hospitals between January 2019 and December 2024. The diagnosis of GLM requires a comprehensive assessment integrating medical history, clinical manifestations, physical examination, imaging, and laboratory findings, with definitive confirmation by histopathological examination. Among these, 575 cases were sourced from a specialized disease registry established by the Department of Breast and Thyroid Surgery at Hospital A in March 2022, which had systematically collected and followed up cases of non-puerperal mastitis since January 2017. The remaining 24 cases were obtained from Hospital B's medical records covering the same period. As of October 1, 2025, the research team extracted eligible cases from Hospital A's registry within the specified timeframe and conducted a unified retrospective review of corresponding cases from Hospital B. According to the predefined exclusion criteria, we excluded 10 patients with missing height information, 31 patients with incomplete lesion diameter records, 8 patients with an unspecified number of lesions, and 8 patients with missing white blood cell (WBC) count data. Furthermore, since the study endpoint was the one-year recurrence rate, an additional 224 patients with less than 12 months of follow-up and no recurrence were excluded. Ultimately, 318 GLM patients with diagnosis dates between January 2019 and December 2024 were included in the final analysis.

The primary endpoint of this study was the one-year recurrence rate, defined as the reappearance of non-puerperal mastitis, ipsilateral or contralateral, within 12 months of the initial diagnosis. Recurrence was defined as the re-emergence of mastitis symptoms following clinical improvement or cure achieved through surgical or conservative treatment. Clinical improvement was characterized by a reduction in lesion size, resolution of skin erythema, significant alleviation of pain, and ultrasonographic evidence of diminished inflammation. Cure was confirmed upon complete resolution of symptoms, absence of palpable masses, and normal findings on both physical and imaging examinations (18). The final follow-up was conducted on September 1, 2025.

In this retrospective design, some collected variables (detailed below) reflected assessments or interventions that occurred during the patient's management course rather than being strictly limited to the baseline state at initial diagnosis. The dataset encompassed multiple clinical variables, including age at diagnosis (years), height (cm), weight (kg), days to first visit (days), defined as the duration from symptom onset to the initial hospital consultation; laterality of the primary lesion (left, right, bilateral); maximum lesion diameter on ultrasound (cm), representing the highest value recorded during the current disease episode; ultrasound-detected lesion count (solitary, multiple); presence of mammary abscess (no, yes); WBC count (10^9/L), indicating the peak measurement observed throughout the clinical course; and documented therapeutic interventions including quinolone therapy (no, yes), penicillin therapy (no, yes), cephalosporin therapy (no, yes), macrolide therapy (no, yes), nitroimidazole therapy (no, yes), antitubercular therapy (no, yes), corticosteroid therapy (no, yes), tetracycline therapy (no, yes), abscess drainage (no, yes), and surgical excision (no, yes). These variables indicated whether a treatment was ever administered, not solely whether it was part of the initial treatment plan.

The study utilized data from two medical institutions under appropriate ethical frameworks. Hospital A's proprietary database was operated in compliance with the Declaration of Helsinki (2013 revision) and received formal approval from its Medical Ethics Committee (approval number: 20220330-024, date: 30.03.2022).

Statistical Analysis

Continuous variables were expressed as mean ± standard deviation, while categorical variables were presented as numbers and percentages. Univariate and multivariate logistic regression (LOG) analyses were performed to identify independent prognostic factors, with odds ratios (ORs) and corresponding 95% confidence intervals (CIs) reported for all significant associations.

The Boruta algorithm, implemented in the Boruta package for R, was a widely used feature selection method based on the random forest (RF) framework (19). It systematically identified all relevant features associated with the prediction target by comparing the importance of original features with randomly generated "shadow features". The primary advantages of the Boruta algorithm included its comprehensiveness, identifying all relevant features rather than identifying only an optimal subset for modeling; robustness, achieved through multiple iterations and statistical testing; and independence from preset parameters, requiring no pre-specified number of features or extensive parameter tuning. In this study, Boruta was first applied to the entire dataset to obtain a stable and interpretable set of predictors. A random seed [set.seed (123)] was set to ensure reproducibility. This fixed feature set was used consistently throughout all subsequent 5-fold cross-validation and model training steps, with no further feature selection performed within the cross-validation loop. This design ensures that all models are trained and compared within the same feature space, facilitating fair performance comparisons and clinical interpretability. To evaluate model performance and mitigate overfitting, a 5-fold cross-validation approach was employed using the caret package. The entire dataset was randomly partitioned into five folds of roughly equal size using a fixed random seed [set.seed (23)]. In each iteration, four folds served as the training subset and the remaining fold served as the validation subset. The median AUC across the five folds was used as the overall performance estimate to select the best-performing algorithm. To provide an intuitive illustration of the model's predictive ability, the fold corresponding to the median AUC (i.e., the centrally located fold representing typical performance) was selected as the representative fold. The model was retrained using the training subset of this representative fold and evaluated on the corresponding validation subset to generate the receiver operating characteristic (ROC) curve and detailed classification metrics (accuracy, sensitivity, specificity, and F1-score). This approach avoids the optimism bias that would result from selecting the best-performing fold, ensuring that the displayed performance reflects the model's typical behavior. Based on the comparative performance evaluation, the best-performing prediction model was deployed as a publicly accessible and free-to-use web calculator through the shiny package (20). This online tool, available free of charge to the research and clinical community, enables real-time recurrence risk prediction based on user-provided clinical features. Furthermore, feature importance ranking was performed on the final model's predictors to identify the most influential variables. The study process is presented in Figure 1. All statistical tests in this study adopted a two-tailed approach, with statistical significance defined at the alpha level of 0.05. The analytical procedures and data visualizations were implemented using R software (version 4.2.2; R Foundation for Statistical Computing, Vienna, Austria).

This study employed five distinct machine learning algorithms, each with its respective implementation. LOG, which estimates the probability of a binary outcome using a logistic function, was implemented with the glm function in the stats package (21). Naïve Bayes (NB), a probabilistic classifier based on Bayes' theorem with feature independence assumption, was performed utilizing the NB function within the e1071 package (22). Linear discriminant analysis (LDA), a method for projecting data into a lower-dimensional space to maximize class separability, was conducted using the LDA function in the MASS package (23). Support vector machine (SVM), which identifies optimal hyperplanes to separate classes with maximum margins, was carried out using the ksvm function from the kernlab package (24). RF, an ensemble technique constructing multiple decision

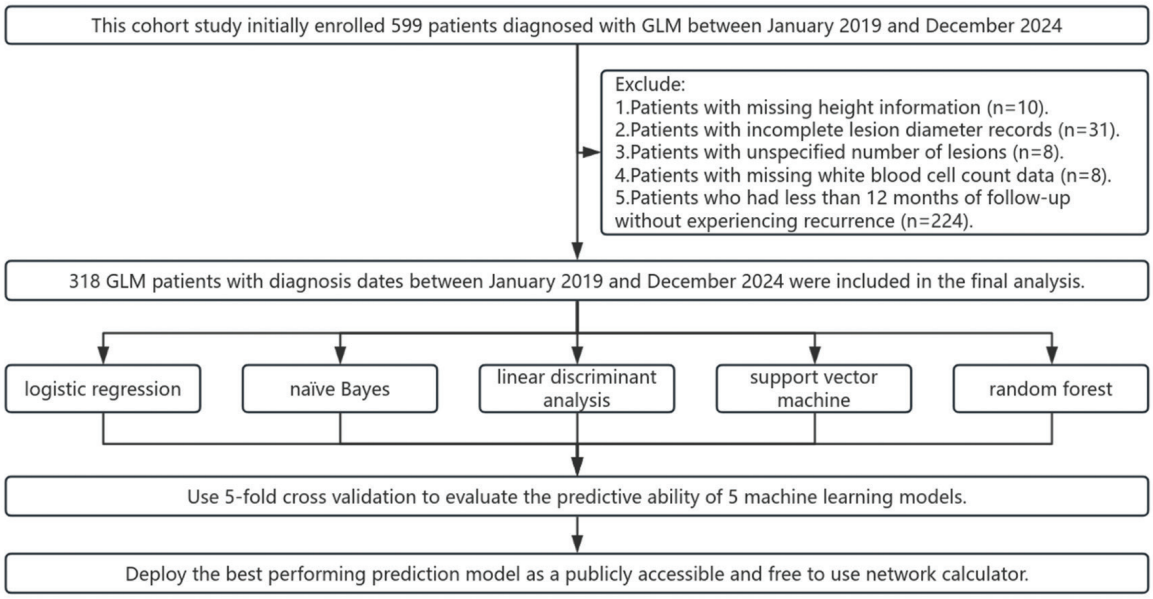

Figure 1. The process of this study
GLM: Granulomatous lobular mastitis

trees to enhance accuracy, was executed with the RF function from the RF package (25). The number of trees (ntree parameter) was set to 183 based on minimizing the out-of-bag (OOB) error rate; to determine this, we trained an initial RF model and identified the point at which the OOB error reached its minimum using the which.min [rf$err.rate (-1)] function. This analysis indicated that 183 trees provided the optimal balance between predictive performance and computational efficiency, as the error rate stabilized beyond this point. A random seed [set.seed (3)] was used for the RF to ensure reproducibility. All models were implemented using their respective packages' default parameters. This approach was adopted to provide a fair and reproducible baseline comparison of algorithmic performance on our dataset, prioritizing generalizability and mitigating the risk of overfitting given the available sample size for model training.

Results

Descriptive Characteristics and Prognostic Factor Analysis

This study included 318 female patients with granulomatous lobular mastitis. With follow-up through September 1, 2025, the cumulative one-year recurrence rate was 32.4%, corresponding to 103 patients who experienced recurrence and 215 who remained recurrence-free. Univariate LOG analysis demonstrated significant associations between GLM recurrence and several clinical factors: origin of the primary lesion ($p = 0.023$), WBC count ($p = 0.001$), antitubercular therapy ($p<0.001$), corticosteroid therapy ($p<0.001$), abscess drainage ($p = 0.002$), and surgical excision ($p = 0.008$) as detailed in Table 1. Multivariate analysis further identified independent predictors of recurrence, including days to first visit ($p = 0.006$; OR: 0.983; 95% CI: 0.971–0.995), origin of primary ($p = 0.007$), antitubercular therapy ($p = 0.002$; OR: 0.358; 95% CI: 0.190–0.676), corticosteroid therapy ($p = 0.005$; OR: 2.990; 95% CI: 1.385–6.452), abscess drainage ($p = 0.003$; OR: 0.290; 95% CI: 0.129–0.653), and surgical excision ($p< 0.001$; OR: 0.190; 95% CI: 0.084–0.428) as presented in Table 1.

Machine Learning

To enhance the prognostic prediction for GLM patients, we employed machine learning approaches, commencing with feature selection using the Boruta package in R (19). The Boruta algorithm, which evaluates feature importance by comparing original attributes with their permuted shadow copies, identified seven significant predictors: age at diagnosis, origin of the primary tumor, WBC count, antitubercular therapy, corticosteroid therapy, abscess drainage, and surgical excision (Figure 2). These seven covariates were incorporated into our subsequent machine learning models.

This study evaluated five machine learning models—LOG, NB, LDA, SVM, and RF—for predicting recurrence in GLM. The AUC values across these models ranged from 0.778 to 0.808, and ROC analysis indicated no statistically significant differences in discriminatory performance among them (Table 2, Figure 3). Among the models, RF achieved the highest F1-score (0.639), accuracy (76.2%), and sensitivity (50%), demonstrating the most balanced performance across multiple metrics. LOG yielded the highest AUC (0.808), while SVM exhibited the highest specificity (95.3%).

Table 1. Univariate and multivariate logistic regression analysis of prognostic factors for 318 patients with GLM

Factors	n (%)/X ± S	Univariate	Multivariate	
		p-value	OR (95% CI)	*p*-value
Age at diagnosis (years)	31.3±4.4	0.189	0.943 (0.883–1.007)	0.080
Height (cm)	159.6±5.1	0.246	1.059 (0.997–1.126)	0.064
Weight (kg)	61.3±11.0	0.546	0.997 (0.970–1.026)	0.850
Days to first visit (days)	26.3±74.9	0.122	0.983 (0.971–0.995)	0.006
Origin of primary		0.023		0.007
Left	164 (51.6)		Ref	
Right	146 (45.9)		1.934 (1.098–3.408)	0.022
Bilateral	8 (2.5)		15.329 (1.755–133.924)	0.014
Maximum lesion diameter on ultrasound (cm)	4.8±2.1	0.272	1.004 (0.989–1.020)	0.600
Ultrasound-detected lesion count		0.539		0.917
Solitary	41 (12.9)		Ref	
Multiple	277 (78.1)		0.951 (0.372–2.431)	
Mammary abscess		0.461		0.413
No	40 (12.6)		Ref	
Yes	278 (87.4)		1.513 (0.561–4.078)	
White blood cell (10^9/L)	10.47±3.98	0.001	1.064 (0.986–1.149)	0.112
Quinolone therapy		0.605		0.578
No	64 (20.1)		Ref	
Yes	254 (79.9)		0.790 (0.344–1.812)	
Penicillin therapy		0.579		0.653
No	306 (96.2)		Ref	
Yes	12 (3.8)		0.713 (0.164–3.108)	
Cephalosporin therapy		0.899		0.450
No	233 (73.3)		Ref	
Yes	85 (26.7)		1.292 (0.665–2.508)	
Macrolide therapy		0.936		0.672
No	303 (95.3)		Ref	
Yes	15 (4.7)		1.326 (0.359–4.888)	
Nitroimidazole therapy		0.786		0.717
No	304 (95.6)		Ref	
Yes	14 (4.4)		1.275 (0.342–4.753)	
Antitubercular therapy		<0.001		0.002
No	190 (59.7)		Ref	
Yes	128 (40.3)		0.358 (0.190–0.676)	
Corticosteroid therapy		<0.001		0.005
No	99 (31.1)		Ref	
Yes	219 (68.9)		2.990 (1.385–6.452)	
Tetracycline therapy		0.653		0.426
No	310 (97.5)		Ref	
Yes	8 (2.5)		0.482 (0.080–2.903)	
Abscess drainage		0.002		0.003
No	88 (27.7)		Ref	
Yes	230 (72.3)		0.290 (0.129–0.653)	
Surgical excision		0.008		<0.001
No	49 (15.4)		Ref	
Yes	269 (84.6)		0.190 (0.084–0.428)	

OR: Odds ratio; CI: Confidence interval; Ref: Reference; GLM: Granulomatous lobular mastitis

Web Application Development

Based on the RF model's balanced performance profile — characterized by the highest F1-score (0.639), accuracy (76.2%), and sensitivity (50%)— this study selected it as the core algorithm for developing a publicly accessible web application (https://w12251393.shinyapps.io/predictGLM/). This decision was guided by several considerations: first, RF demonstrated the most balanced performance across key classification metrics, with a high F1-score reflecting the optimal balance between precision and recall, and its highest sensitivity indicating an enhanced ability to identify true recurrence cases, which is particularly crucial for clinical early warning; second, as an ensemble learning algorithm, RF exhibited greater generalization and robustness, enabling better adaptation to new data and making it suitable for deployment as the core prediction engine in a public application. All five models achieved comparable discriminatory performance, with no statistically significant differences in AUC, and the selection of RF represents a pragmatic choice based on its multi-metric balance rather than a claim of statistical superiority. The platform automatically calculates the one-year recurrence risk based on patient characteristics entered by the user. Figure 4 displays a functional example of the interface of this web application.

Based on the RF model, feature importance ranking for the seven covariates was performed using the importance function from the RF package in R, which employs a permutation-based approach to evaluate variable significance by measuring the mean decrease in accuracy when out-of-bag data for each predictor is randomly shuffled (25). The analysis revealed the following descending order of predictive importance: WBC count, which emerged as the most influential predictor, followed by age at diagnosis, origin of the primary, surgical excision, antitubercular therapy, corticosteroid therapy, and abscess drainage (Figure 5).

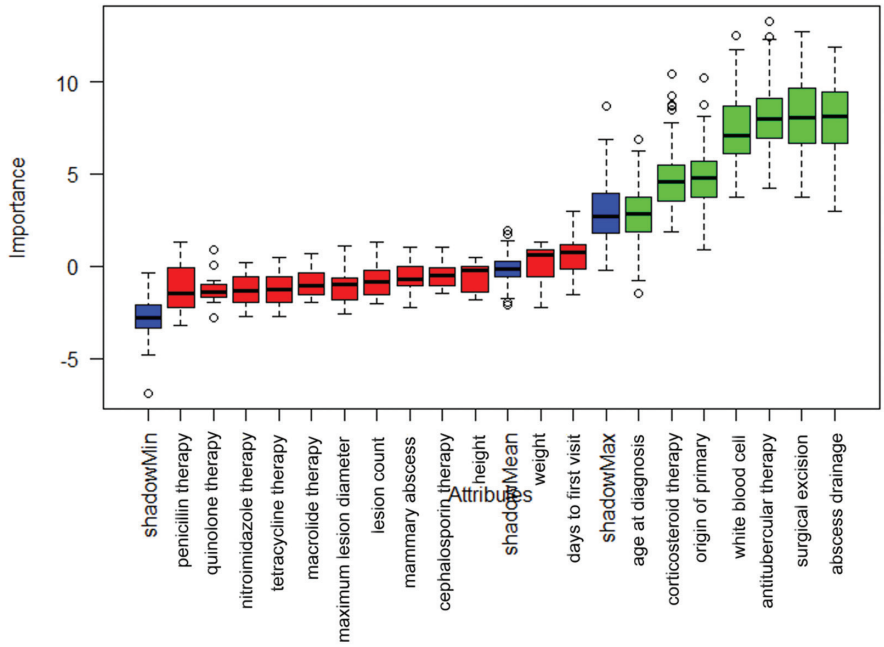

Figure 2. Predictor selection for machine learning modeling

Table 2. Performance comparison of the five machine learning models for GLM recurrence prediction in the validation set

Mode	Accuracy (%)	AUC	F1-score	Sensitivity (%)	Specificity (%)	$P_{vs\ LOG}$	$P_{vs\ NB}$	$P_{vs\ LDA}$	$P_{vs\ SVM}$
LOG	69.8	0.808	0.541	40.0	83.7	Ref	0.7129	0.9177	0.6496
NB	70.3	0.778	0.435	28.6	90.7	0.7129	Ref	0.712	0.9147
LDA	69.8	0.807	0.536	40.0	81.4	0.9177	0.712	Ref	0.662
SVM	74.6	0.787	0.456	30.0	95.3	0.6496	0.9147	0.662	Ref
RF	76.2	0.785	0.639	50.0	88.4	0.7776	0.9271	0.7868	0.9837

AUC: Area under the curve; LOG: Logistic regression; Ref: Reference; NB: Naïve bayes; LDA: Linear discriminant analysis; SVM: Support vector machine; RF: Random forest; GLM: Granulomatous lobular mastitis

Discussion and Conclusion

Studies have reported that the recurrence rate of GLM can be as high as 24–40%, making it a commonly recurring breast condition (8-11). Accurate prediction of recurrence could inform treatment decisions and follow-up strategies, ultimately improving patient outcomes. Although some studies have attempted to use staging systems to predict patient prognosis (12, 14, 15), these systems have generally demonstrated limited predictive efficacy due to the substantial heterogeneity of the condition and the wide variation in clinical manifestations among individual patients.

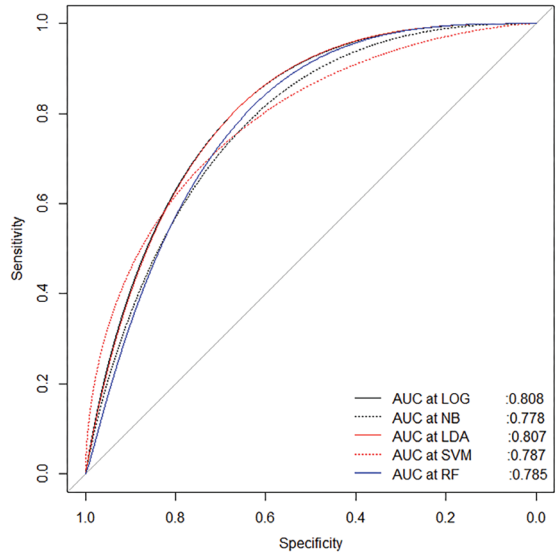

Figure 3. Comparison of ROC curves for the five machine learning models
AUC: Area under the curve; LOG: Logistic regression; NB: Naïve bayes; LDA: Linear discriminant analysis; SVM: Support vector machine; RF: Random forest; ROC: Receiver operating characteristic

Using a multicenter retrospective cohort, this study systematically compared five machine learning models for predicting one-year recurrence risk in GLM and subsequently developed a publicly accessible online calculator based on the optimally performing RF model. The results demonstrated that all models achieved comparable discriminatory performance, with AUCs ranging from 0.778 to 0.808. The RF model exhibited a balanced performance profile with an F1-score of 0.639, accuracy of 76.2%, and sensitivity of 50%, while LOG achieved the highest AUC (0.808) and the SVM exhibited the highest specificity (95.3%). Based on its balanced multi-metric performance and inherent feature importance interpretation, RF was selected as the final model for clinical deployment; this choice reflects practical considerations rather than statistical superiority, given the equivalent AUCs across models. Feature importance analysis identified WBC as the most influential predictor of recurrence, followed by age at diagnosis, origin of primary, surgical excision, antitubercular therapy, corticosteroid therapy, and abscess drainage. These findings provide novel insights and a practical tool for individualized recurrence risk assessment in GLM.

The predictive performance of our models aligns closely with previously reported machine learning applications in GLM.

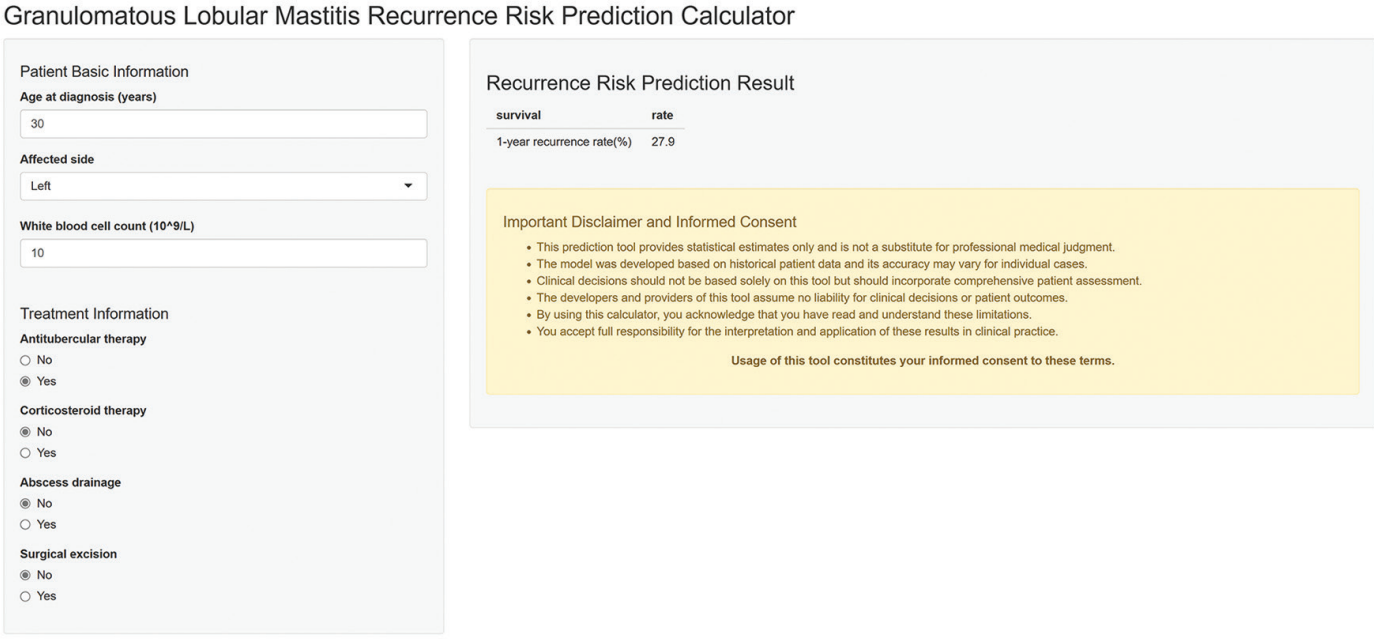

Figure 4. An example showing web function (https://w12251393.shinyapps.io/predictGLM/)

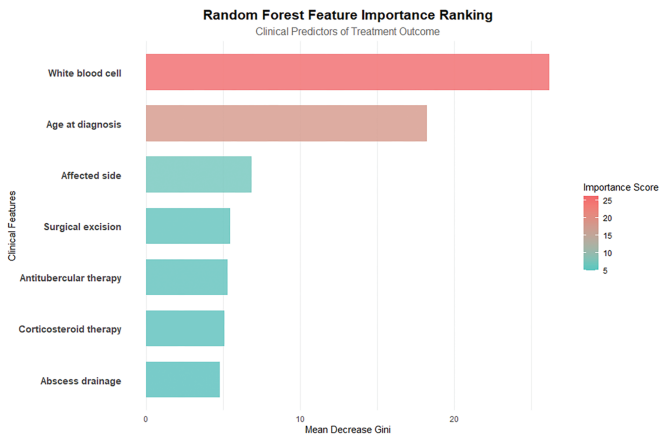

Figure 5. Feature importance ranking in the random forest model

Li et al. (16) analyzed 212 GLM patients and compared LOG, RF, and neural network models, reporting that the neural network outperformed the other models in their specific dataset, while their RF model yielded an AUC of 0.793—remarkably similar to the 0.785 AUC observed in our RF model. Similarly, Ma et al. (17) developed an XGBoost model based on contrast-enhanced ultrasound features and reported an AUC of 0.808, which is also comparable to our findings. These convergent results across studies and populations suggest that current machine learning approaches for predicting GLM recurrence consistently achieve AUCs of 0.78–0.81, reflecting the inherent complexity of GLM recurrence.

Despite comparable predictive performance, our study extends previous work in several important aspects. First, our models were developed and validated in a multicenter cohort (n = 318), enhancing generalizability compared with single-center studies with smaller samples. Second, we systematically compared five algorithms under standardized conditions, thereby providing a comprehensive benchmark for future research. Third, and most significantly, we implemented the optimal model as a freely accessible web-based calculator (https://w12251393.shinyapps.io/predictGLM/), filling a critical gap in clinical implementation that previous studies had not addressed.

The clinical deployment of any prediction model requires careful consideration of its performance characteristics. Our RF model's sensitivity of 50% warrants attention, as it indicates that approximately half of patients who ultimately experience recurrence may not be identified by the model (false negatives). In clinical practice, such false-negative predictions could lead to inadequate monitoring intensity or delayed therapeutic interventions, potentially compromising patient outcomes. Therefore, it should be emphasized that this web-based calculator is intended as a risk-stratification aid rather than a definitive predictive instrument. Clinicians should integrate model predictions with comprehensive clinical evaluation, maintaining standard follow-up protocols for all patients and exercising heightened vigilance for those with strong clinical suspicion of recurrence despite low-risk model predictions. Conversely, the model's relatively high specificity (88.4%) offers meaningful clinical value by reliably identifying patients at low risk of recurrence. This capability may facilitate more efficient resource allocation, potentially reducing unnecessarily frequent follow-up visits or overly aggressive treatment among low-risk populations. From a clinical utility perspective, the tool may be better suited to rule out low-risk patients than to definitively identify high-risk individuals—a distinction that should guide its appropriate integration into clinical workflows.

The identification of WBC as the most important predictor in our RF model warrants careful interpretation. While Sun et al. (26) similarly reported WBC as the predominant risk factor for GLM recurrence, several other studies found no significant association between WBC and recurrence (11, 27, 28). Several factors may explain these discrepancies. First, the timing of WBC measurement varied substantially across studies: our study used peak WBC during the disease course, whereas others employed baseline WBC at diagnosis or random measurements of WBC. Given that WBC can fluctuate in response to disease activity and therapeutic interventions, measurement timing critically influences its prognostic value. Second, treatment-related factors (antibiotics, corticosteroids) can directly modulate WBC levels, potentially confounding the relationship between WBC and recurrence in retrospective analyses. Third, and perhaps most importantly, the etiological heterogeneity of GLM likely influences both WBC patterns and recurrence risk. For instance, infectious etiologies (e.g., tuberculosis, Corynebacterium infection) may exhibit distinct inflammatory profiles compared with idiopathic or autoimmune variants, yet our inability to perform etiological stratification may have averaged out these differences.

The "days to first visit" variable exhibited substantial variability (mean 26.3±74.9 days), reflecting a right-skewed distribution due to a subset of patients with markedly delayed presentation. Its modest protective effect (OR: 0.983) should be interpreted cautiously, as it may reflect confounding by disease severity—patients with milder symptoms may both delay care and have a lower recurrence risk—rather than a direct causal relationship.

The inclusion of treatment-related variables—surgical excision, corticosteroid therapy, antitubercular therapy, and abscess drainage—in the final model reflects their prognostic significance. LOG revealed strong associations between recurrence and surgical excision, corticosteroid therapy, and antitubercular therapy.

These findings should be interpreted with caution due to potential confounding by indication, a common issue in observational studies where treatment assignment is influenced by disease severity. For instance, patients receiving corticosteroids may have more severe inflammation and therefore a higher risk of recurrence regardless of treatment effect. Conversely, the protective effects of surgery and antitubercular therapy may reflect patient selection rather than causal benefits. These variables may also indirectly capture etiological information—e.g., antitubercular therapy suggests tuberculous mastitis, which is characterized by distinct recurrence patterns. However, without systematic etiological confirmation, such inferences remain speculative. Therefore, these associations should be understood as observational prognostic factors and not as evidence of causal treatment effects. Prospective studies with standardized protocols and etiological stratification are needed to establish causality.

Despite the encouraging results, our study was constrained by several limitations. First, the retrospective design inherently carried a risk of selection bias. To accurately assess the one-year recurrence endpoint and avoid outcome misclassification, we excluded patients with <12 months of follow-up who had no recurrence (n = 224). While methodologically necessary, this might have limited the cohort's representativeness, as patients lost to follow-up could have differed systematically. Notably, the one-year endpoint itself represented only one aspect of this potential bias. This limitation might have affected the model's generalizability to broader populations. Second, while we evaluated several machine learning models based on structured clinical data, we did not investigate deep learning approaches or incorporate imaging-derived features (e.g., radiomic or ultrasound-derived features), which might have captured more complex patterns in the data, though at the cost of interpretability. Our deliberate focus on readily available clinical variables was intended to maximize clinical applicability and ease of implementation; however, this choice meant that potentially informative imaging data were not utilized. Future studies integrating multi-modal data may further improve predictive accuracy. Third, despite our multicenter dataset (two centers, n = 318), this study lacks external validation in an independent cohort. All data were utilized for model development and internal validation; therefore, the model's performance on entirely unseen populations remains unknown. This represents a critical limitation, as the current results may not fully reflect the model's generalizability to broader clinical settings. Future research should prioritize external validation using independent cohorts with similar or larger sample sizes. We are actively seeking collaborations with additional centers to prospectively collect validation data, which will be essential before the model can be considered for wider clinical implementation. Fifth, the lack of etiological subtyping represents an important limitation. GLM encompasses a spectrum of disorders with diverse etiologies (e.g., infectious, autoimmune, idiopathic) that have substantially different treatment responses and recurrence patterns. Due to the retrospective design and the absence of standardized etiological screening (including testing for tuberculosis, Corynebacterium, fungi, and autoimmune antibodies) in routine clinical practice, we were unable to stratify patients according to these subtypes. Consequently, our machine learning models reflect average effects across a mixed population, and heterogeneity among subtypes may have diluted certain predictive signals. Sixth, while our study focused on discrimination metrics, calibration assessment—an important aspect of model performance for clinical decision-making—was not performed. Future external validation studies should include comprehensive calibration evaluation, including calibration curves and Brier scores, to further establish the model's clinical utility.

However, this study established a multicenter, large-scale cohort and employed multiple machine learning algorithms to predict recurrence risk and subsequently developed a clinically applicable web-based calculator. This tool incorporates influential features such as treatment modalities and offers significant clinical value by supporting personalized recurrence risk assessment and treatment decision-making. Future research should aim to enhance the model's clinical utility further. Specifically, prospective studies should incorporate standardized etiological screening protocols and develop dedicated predictive models for different subtypes—especially those requiring differentiation from specific infections—to enable precision diagnosis and treatment. Key directions include improving sensitivity to reduce false-negative predictions. Potential strategies encompass prospectively collecting early or dynamic biomarkers, employing advanced machine learning techniques to handle imbalanced data, and expanding multi-center collaborations to enrich the sample size, particularly within the recurrence subgroup, thereby refining the model's ability to identify recurrence patterns.

Study Limitations

Using a multicenter cohort, this study successfully developed and validated a RF-based prediction model for GLM recurrence risk; the model was selected for its balanced performance across multiple metrics and subsequently translated into a clinically accessible web-based calculator. All five machine learning models demonstrated comparable discriminatory performance, with no statistically significant differences in AUC. The tool is intended as a risk stratification aid to support clinical decision-making, not as a definitive predictive instrument. Although the retrospective design and the absence of deep learning models represented limitations, this work provided a practical tool for personalized prognostic assessment by integrating key clinical features. External validation on larger and more diverse populations was recommended to further enhance the model's clinical utility.

Ethics

Ethics Committee Approval: The study utilized data from two medical institutions under appropriate ethical frameworks. Hospital A's proprietary database was operated in compliance with the Declaration of Helsinki (2013 revision) and received formal approval from its Medical Ethics Committee (approval number: 20220330-024, date: 30.03.2022).

Informed Consent: Retrospective study.

Footnotes

Authorship Contributions

Surgical and Medical Practices: J.F., R.W., J.C., Y.L.; Concept: J.F., R.W.; Data Collection or Processing: J.F., R.W., J.C.; Analysis or Interpretation: J.F., J.C.; Literature Search: J.F., Y.L.; Writing: J.F., R.W., J.C., Y.L.

Conflict of Interest: No conflict of interest was declared by the authors.

Financial Disclosure: The authors declared that this study received no financial support.

References

1. Ma X, Min X, Yao C. Different treatments for granulomatous lobular mastitis: a systematic review and meta-analysis. Breast Care (Basel). 2020; 15: 60-66. (PMID: 32231499) [Crossref]
2. Farouk O, Abdelkhalek M, Abdallah A, Shata A, Senbel A, Attia E, et al. Rifampicin for idiopathic granulomatous lobular mastitis: a promising alternative for treatment. World J Surg. 2017; 41: 1313-1321. (PMID: 28050664) [Crossref]
3. Li Y, Chen L, Zhang C, Wang Y, Hu J, Zhou M, et al. Clinicopathologic features and pathogens of granulomatous lobular mastitis. Breast Care (Basel). 2023; 18: 130-140. (PMID: 37261131) [Crossref]
4. Liu R, Luo Z, Dai C, Wei Y, Yan S, Kuang X, et al. Corynebacterium parakroppenstedtii secretes a novel glycolipid to promote the development of granulomatous lobular mastitis. Signal Transduct Target Ther. 2024; 9: 292. (PMID: 39428541) [Crossref]
5. Zeng W, Lao S, Jia W, Shen X, Wu L, Zhong Y, et al. Clinical features and recurrence of Corynebacterium kroppenstedtii infection in patients with mastitis. BMC Womens Health. 2022; 22: 276. (PMID: 35794560) [Crossref]
6. Wang J, Zhang Y, Lu X, Xi C, Yu K, Gao R, et al. Idiopathic granulomatous mastitis with skin rupture: a retrospective cohort study of 200 patients who underwent surgical and nonsurgical treatment. J Invest Surg. 2021; 34: 810-815. (PMID: 31818161) [Crossref]
7. Çetin K, Sıkar HE, Göret NE, Rona G, Barışık NÖ, Küçük HF, et al. Comparison of topical, systemic, and combined therapy with steroids on idiopathic granulomatous mastitis: a prospective randomized study. World J Surg. 2019; 43: 2865-2873. (PMID: 31297582) [Crossref]
8. Çetin K, Sıkar HE, Feratoğlu F, Taşdoğan B, Güllüoğlu BM. Treatment of granulomatous mastitis with steroids: should the decision to end the treatment be made radiologically? Eur J Breast Health. 2023; 20: 25-30. (PMID: 38187102) [Crossref]
9. Li S, Huang Q, Song P, Han X, Liu Z, Zhou L, et al. Clinical characteristics and therapeutic strategy of granulomatous mastitis accompanied by Corynebacterium kroppenstedtii: a retrospective cohort study. BMC Womens Health. 2023; 23: 388. (PMID: 37491234) [Crossref]
10. Karanlik H, Ozgur I, Simsek S, Fathalizadeh A, Tukenmez M, Sahin D, et al. Can steroids plus surgery become a first-line treatment of idiopathic granulomatous mastitis? Breast Care (Basel). 2014; 9: 338-342. (PMID: 25759614) [Crossref]
11. Li Q, Wan J, Feng Z, Shi J, Wei W. Predictive significance of the preoperative neutrophil-lymphocyte ratio for recurrence in idiopathic granulomatous mastitis patients. Am Surg. 2023; 89: 5577-5583. (PMID: 36880848) [Crossref]
12. Wu R, Zhang H, Wang Y, Mo Y, Hu H, Chen J, et al. A new stage for predicting the prognosis of granulomatous lobular mastitis. PLoS One. 2025; 20: e0319956. (PMID: 40106498) [Crossref]
13. Basim P, Argun D, Argun F. Risk factors for idiopathic granulomatous mastitis recurrence after patient-tailored treatment: do we need an escalating treatment algorithm? Breast Care (Basel). 2022; 17: 172-179. (PMID: 35707181) [Crossref]
14. Yılmaz TU, Gürel B, Güler SA, Baran MA, Erşan B, Duman S, et al. Scoring idiopathic granulomatous mastitis: an effective system for predicting recurrence? Eur J Breast Health. 2018; 14: 112-116. (PMID: 29774320) [Crossref]
15. Yaghan R, Hamouri S, Ayoub NM, Yaghan L, Mazahreh T. A proposal of a clinically based classification for idiopathic granulomatous mastitis. Asian Pac J Cancer Prev. 2019; 20: 929-934. (PMID: 30912417) [Crossref]
16. Li L, Yang W, Jia H. Deep learning models for predicting the recurrence of idiopathic granulomatous mastitis. J Inflamm Res. 2025; 18: 2943-2953. (PMID: 40026307) [Crossref]
17. Ma L, Du P, Sun X, Zhu L, Li Y, Li X, et al. Correlation analysis and construction of a predictive model between contrast-enhanced ultrasound features and the risk of recurrence in granulomatous mastitis. Acad Radiol. 2025; 32: 3170-3180. (PMID: 39843281) [Crossref]
18. Zhou F, Liu L, Yu ZG. Expert consensus on the diagnosis and treatment of non-puerperal mastitis. Chinese Journal of Practical Surgery. 2016;36:755-758. [Crossref]
19. Kursa MB, Rudnicki WR. Feature Selection with the Boruta Package. Journal of Statistical Software. 2010; 36: 1-13. [Crossref]
20. Chang W, Cheng J, Allaire J, Sievert C, Schloerke B, Xie Y, et al. Shiny: web application framework for R. 2022. R package version 1.7.4. Available link: https://CRAN.R-project.org/package=shiny. [Crossref]
21. R Core Team. R: a language and environment for statistical computing. 2022. R Foundation for Statistical Computing, Vienna, Austria. Available link: https://www.R-project.org/. [Crossref]
22. Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F. e1071: misc functions of the department of statistics, probability theory group (Formerly: E1071), TU Wien. 2023. R package version 1.7-14. Available link: https://CRAN.R-project.org/package=e1071. [Crossref]
23. Venables WN, Ripley BD. Modern applied statistics with S. Fourth Edition. Springer, New York; 2022.
24. Karatzoglou A, Smola A, Hornik K. Kernlab: kernel-based machine learning lab. 2023. R package version 0.9-32. Available link: https://CRAN.R-project.org/package=kernlab. [Crossref]
25. Liaw A, Wiener M. Classification and regression by randomforest. R News 2002; 2: 18-22. Available link: https://journal.r-project.org/articles/RN-2002-022/RN-2002-022.pdf [Crossref]
26. Sun J, Shao S, Wan H, Wu X, Feng J, Gao Q, et al. Prediction models for postoperative recurrence of non-lactating mastitis based on machine learning. BMC Med Inform Decis Mak. 2024; 24: 106. (PMID: 38649879) [Crossref]

27. Velidedeoglu M, Kundaktepe BP, Aksan H, Uzun H. Preoperative fibrinogen and hematological indexes in the differential diagnosis of idiopathic granulomatous mastitis and breast cancer. Medicina (Kaunas). 2021; 57: 698. (PMID: 34356979) [Crossref]

28. Ciftci AB, Bük ÖF, Yemez K, Polat S, Yazıcıoğlu İM. Risk factors and the role of the albumin-to-globulin ratio in predicting recurrence among patients with idiopathic granulomatous mastitis. J Inflamm Res. 2022; 15: 5401-5412. (PMID: 36158516) [Crossref]