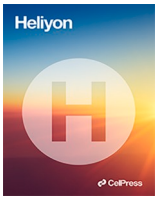
Heliyon 10 (2024) e31307

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Research article

Predicting the risk of primary Sjögren's syndrome with key N7-methylguanosine-related genes: A novel XGBoost model

Hui Xie [a,b,1], Yin-mei Deng [c,1], Jiao-yan Li [d], Kai-hong Xie [e], Tan Tao [b,*], Jian-fang Zhang [f,**]

[a] Department of Radiotherapy, Affiliated Hospital (Clinical College) of Xiangnan University, Chenzhou, 423000, PR China
[b] Faulty of Applied Sciences, Macao Polytechnic University, Macao, 999078, PR China
[c] Department of Nursing, Affiliated Hospital (Clinical College) of Xiangnan University, Chenzhou, 423000, PR China
[d] Department of Rheumatology and Clinical Immunology, The First Hospital of Changsha, 410005, Changsha, PR China
[e] Department of Oncology, Affiliated Hospital (Clinical College) of Xiangnan University, Chenzhou, 423000, PR China
[f] Department of Physical Examination, Center for Disease Control and Prevention of Beihu District, Chenzhou, 423000, PR China

ARTICLE INFO

Keywords:
N7-methylguanosine
Machine learning
Primary Sjögren's syndrome
Gene expression omnibus database

ABSTRACT

Objectives: N7-methylguanosine (m7G) plays a crucial role in mRNA metabolism and other biological processes. However, its regulators' function in Primary Sjögren's Syndrome (PSS) remains enigmatic.
Methods: We screened five key m7G-related genes across multiple datasets, leveraging statistical and machine learning computations. Based on these genes, we developed a prediction model employing the extreme gradient boosting decision tree (XGBoost) method to assess PSS risk. Immune infiltration in PSS samples was analyzed using the ssGSEA method, revealing the immune landscape of PSS patients.
Results: The XGBoost model exhibited high accuracy, AUC, sensitivity, and specificity in both training, test sets and extra-test set. The decision curve confirmed its clinical utility. Our findings suggest that m7G methylation might contribute to PSS pathogenesis through immune modulation.
Conclusions: m7G regulators play an important role in the development of PSS. Our study of m7G-realted genes may inform future immunotherapy strategies for PSS.

1. Introduction

Primary Sjogren's syndrome (PSS), a prevalent chronic autoimmune disease, predominantly affects women with a prevalence ranging from 0.5 % to 1.0 % [1]. The main target of PSS is the exocrine glands, including the lacrimal and salivary glands, but it can also impact the respiratory, blood, nervous, and other bodily systems, posing a substantial threat to the patients' lives [2]. PSS is categorized as an autoimmune disorder, and it is reported that approximately 30 % of patients suffering from other autoimmune

* Corresponding author. Faulty of Applied Sciences, Macao Polytechnic University, Macao, 999078, PR China.
** Corresponding author. Department of Physical Examination, Center for Disease Control and Prevention of Beihu District, Chenzhou, 423000, Hunan province, PR China.
E-mail addresses: taotanjs@gmail.com (T. Tao), czzhangjf168@163.com (J.-f. Zhang).
[1] Hui Xie, and Yin-mei Deng contributed equally to this study.

https://doi.org/10.1016/j.heliyon.2024.e31307
Received 22 November 2023; Received in revised form 10 May 2024; Accepted 14 May 2024
Available online 16 May 2024

diseases also have PSS [3]. However, due to the absence of specific diagnostic markers, diagnosis is often delayed, resulting in a lag of approximately 6–10 years between initial symptoms and confirmation [4]. Immune dysfunction has been identified as a crucial factor in the pathogenesis of PSS, with mild lesions characterized by local infiltration of $CD4^+$ and $CD8^+$ T cells, and moderate to severe lesions dominated by the production of autoantibodies from B cells [5]. Overactivation of B cells can lead to salivary gland abnormalities and increased lymphoma risk. Given the clinical heterogeneity and complex pathological mechanisms underlying PSS, elucidating its underlying mechanisms remains unclear. Therefore, identifying relevant biomarkers and immune molecules is crucial for understanding its intricate immune processes.

Increasing evidence suggests that the occurrence and progression of PSS are determined not only by genetic variations but also by epigenetic dysregulation [6,7]. RNA modifications, as crucial components of epigenetic modifications, play a pivotal role in regulating various physiological processes and disease pathogenesis [8]. Notably, the dynamic regulation and dysregulation of these RNA modifications are intimately associated with the initiation, maintenance, and progression of PSS [9]. Among the numerous RNA modifications, N6-methyladenosine (m6A), 5-methylcytosine (m5C), and N7-methylguanosine (m7G) are particularly common [10]. Crucially, m7G is the most frequent RNA cap modification found in various RNAs of eukaryotic organisms, exerting profound effects on RNA metabolism, processing, and function [11]. Due to technological constraints, the study of m7G-related regulatory factors in diseases like PSS has only recently gained attention. Research has shown that METTL1 and WDR4, overexpressed in nasopharyngeal carcinoma (NPC) and associated with poor outcomes, can be inhibited to suppress tumor growth, migration, and invasion [12]. Additionally, they influence the immunomodulatory tumor microenvironment, regulating immune cell infiltration and

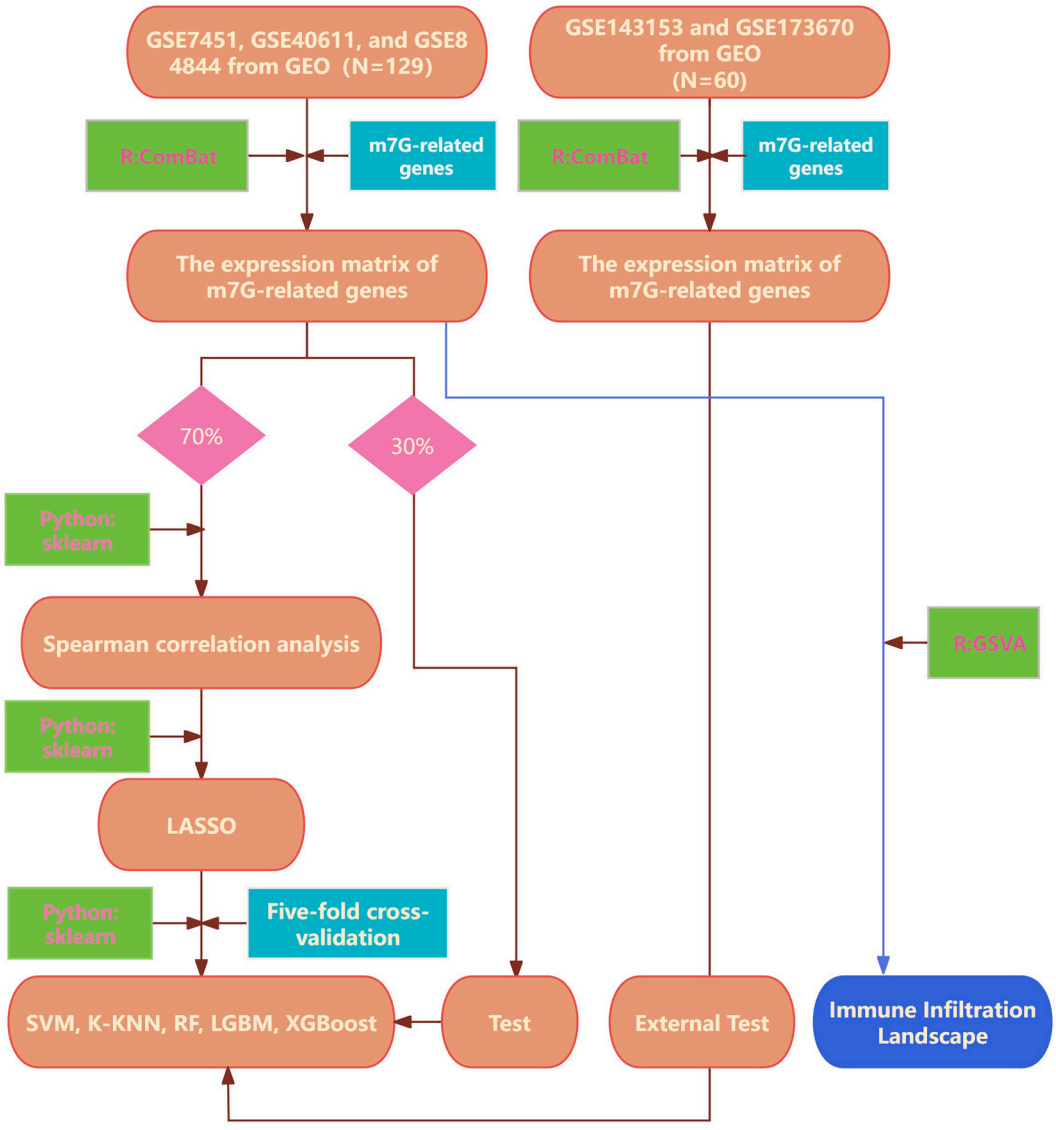


Fig. 1. The overall technology flowchart.

cancer-immune cell interactions [13]. The latest research conducted by Liu et al. has uncovered the crucial role of m7G-related genes in the pathogenesis and development of PSS [14]. Despite this progress, the relationship between m7G-associated regulatory factors and the immune regulation of PSS remains elusive. Therefore, it is imperative to intensify research efforts in this area to gain a deeper understanding of the complex mechanisms underlying PSS and potentially identify novel therapeutic targets.

The field of machine learning (ML) has revolutionized the way we can derive meaningful patterns and insights from complex, large-scale datasets. Its successful implementation is evident across various sectors, including data mining, pattern recognition, and bioinformatics. In this context, our research seeks to delve into gene expression data related to PSS, sourced from the publicly available Gene Expression Omnibus (GEO) database. By integrating bioinformatics tools with ML algorithms, we are committed to evaluating the predictive capacity of m7G-related genes in the onset of PSS. Our overarching goal is to uncover new avenues for clinical diagnosis, treatment, and prognosis of PSS, thereby contributing to significant advancements in this area of study.

Furthermore, our groundbreaking approach marks the first instance of exploring m7G-related genes in PSS, a territory that has remained largely unexplored. By harnessing the power of an interdisciplinary research approach that merges the sophistication of bioinformatics with the precision of machine learning, we aim to unlock profound insights into PSS that traditional methods may have overlooked. This innovative quest offers us a fresh perspective and holds the potential to deepen our understanding of PSS, thereby paving the way for future breakthroughs.

2. Materials and methods

The research methodology is illustrated in Fig. 1.

2.1. Data collection

In our research, we amassed 189 samples from five GEO datasets (http://www.ncbi.nlm.nih.gov/geo): GSE7451 (10 normal and 10 PSS saliva samples), GSE40611 (18 normal and 31 PSS parotid gland samples), GSE84844 (30 normal and 30 PSS blood samples), GSE143153 (7 normal and 25 PSS parotid gland samples), and GSE173670 again (14 normal and 14 blood saliva samples).The integrated data from chips GSE7451, GSE40611, and GSE84844 within the R environment were used to establish a predictive model and served as the primary dataset (Data SetI) for analysis in this study. Similarly, datasets GSE143153 and GSE173670 were integrated in the R environment to serve as an independent external test set (Data Set II) for the predictive model. Each chip data set has undergone normalization of gene expression levels using the reads per kilobase per million mapped reads (RPKM) method on the raw data. The batch effect was successfully mitigated for both sets of data using the "ComBat" function from the "SVA" package) [15] in R software (4.0.2, http://www.r-project.org). This approach effectively eliminated errors caused by different chip sequencing while preserving the variations between individuals.

2.2. Extraction of m7G-related gene expression matrix

Within the R environment, we successfully isolated the expression matrix of m7G-related genes. Prior to this, we conducted a thorough examination of pertinent literature [16] and accessed databases such as GSEA (gsea-msigdb.org) to gather relevant information for our investigation. Consequently, we recognized a sum of 29 genes connected with m7G regulation. The extracted m7G-related gene expression matrix was normalized using the Z-score method, and all subsequent studies were conducted based on this processed matrix. The exhaustive compilation of these genes is presented in Supplementary Table 1.

2.3. Screening of key genes

Across the entire cohort of samples included in the study, a labeling system was employed where normal samples were designated as 0 and PSS samples were labeled as 1. The data *Set*I were then randomly apportioned into training and test sets, adhering to a 7:3 ratio. Within the Python environment, a Spearman correlation analysis was conducted on the training set data using the spearman function from the scipy library (https://github.com/scipy/scipy/, accessed on November 1, 2021) [16]. Whenever the correlation between two genes exceeded 0.9, one of the pair was randomly removed, ensuring that only one gene from each highly correlated pair advanced to the next stage of analysis [17,18]. In the Python environment, the Least Absolute Shrinkage and Selection Operator (LASSO) analysis was implemented as the final selection process for identifying the key genes. This was accomplished using the Lasso class from the sklearn (Scikit-learn) library [19]. LASSO is a method used in data mining that helps simplify and optimize regression models. It does this by adding a penalty to the regular linear regression model, encouraging some of the coefficients to shrink towards zero, and sometimes even become exactly zero. This process reduces model complexity, avoids overfitting, and can even help identify the most important variables in the dataset [20]. In essence, LASSO helps create more robust and interpretable models by balancing the fit to the data with the simplicity of the model. To conduct a more thorough evaluation of the impact of the selected key genes on PSS risk, this study has further employed a binary logistic regression model to assess the significance of the coefficients of these genes with regard to PSS. Utilizing the LASSO method, we are able to calculate the PSS risk score, which provides a quantitative measure of the PSS's risk level based on their specific characteristics and features. This risk score is derived from the LASSO model, which allows us to identify and assign weights to the most predictive variables in predicting a particular outcome or response. By calculating the PSS risk score, we can gain a deeper understanding of the risk factors associated with PSS, enabling more informed decisions and targeted interventions in patient of PSS care and management. Furthermore, we have generated a heatmap depicting the expression patterns of

risk genes, illustrating the differential expression between normal and PSS states.

The Risk_score of the model can be calculated as follows:

$$Risk\_Score = \sum_{i} \text{feature}_i \times \text{Coefficients} + b \quad (1)$$

Coefficient is obtained by an iterative of LASSO algorithm,i represents the gene,and b represents the intercept term.

In Equation (1), the intercept term "b" is obtained through LASSO regression using the scikit-learn library in a Python environment [19,21]. To maximize the stability of the regression coefficients and intercept terms for key genes, we fixed the crucial regularization parameter λ determined during the first convergence of LASSO regression. Subsequently, we carefully designed multiple sets of experiments by adjusting the random seed, the number of folds in cross-validation, and randomly increasing and decreasing the sample size of the dataset to simulate diverse experimental conditions. These experiments aimed to obtain the average values of the regression coefficients and bias terms for key genes under different conditions. Ultimately, we will use these average values as core parameters to update the regression coefficients and bias terms in Formula 1, ensuring that the model maintains excellent stability and accuracy in applications.

2.4. Construction of the prediction model

For devising the predictive model for PSS in this study, a variety of machine learning algorithms were deployed, including Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Random Forest (RF), Light Gradient Boosting Machine (LightGBM), and Extreme Gradient Boosting Decision Tree (XGBoost). The modeling process is entirely carried out in Python software (3.8.5 , https://www.python.org/downloads/release/python-380/). For KNN, we're using KNeighborsClassifier from sklearn.neighbors [19], which assigns labels to new instances based on their proximity to similar training instances. In the KNN algorithm, we set the weights to distance, which means that during prediction, each neighbor's voting weight is inversely proportional to its distance from the query point, i.e., closer neighbors will have a greater influence. To find the optimal setting for n_neighbors, we chose the GridSearchCV strategy. This strategy enables us to thoroughly and systematically test different values for n_neighbors to determine which one results in the best performance for the model on the test set. After conducting the search, this study ultimately determined that the optimal value for n_neighbors is 5. The RF classifier, implemented with RandomForestClassifier from sklearn.ensemble [19], combines multiple decision trees to improve classification accuracy. To ensure the optimal performance of the classifier, we have set criterion = 'gini' as the splitting criterion. By leveraging GridSearchCV, we systematically evaluate various combinations of hyperparameters to precisely identify the optimal values for each. This rigorous approach guarantees that the random forest classifier operates at its utmost potential, utilizing the ideal number of trees (n_estimators = 386), tree depth (max_depth = 6), and splitting criteria (min_impurity_decrease = 0.0003). Consequently, the overall performance of the classifier is significantly enhanced. SVM is handled by the SVC model from sklearn.svm [22]. The parameters are tuned with a convergence measure of 0.1, a radial basis function (rbf) kernel, and a regularization factor of 0.5. LightGBM is incorporated through its LGBMClassifier from the lightgbm package [23], configured for fast and accurate gradient boosting. To optimize the performance of the model, we also utilized the GridSearchCV strategy to systematically search for the best combination of hyperparameters. These hyperparameters include n_estimators, max_depth, the maximum number of leaves per tree (num_leaves), and the learning_rate. After thorough searching, we have identified the following optimal parameter values: n_estimators = 56, max_depth = 6, num_leaves = 29, and learning_rate = 0.05. These parameter configurations will further enhance the model's predictive capabilities and generalization ability. XGBoost is employed via XGBClassifier from the xgboost library [24], another gradient boosting framework particularly effective with large datasets. To discover the optimal combination of parameters, we utilize GridSearchCV to systematically tune the crucial hyperparameters of XGBoost: n_estimators, max_depth, learning_rate, colsample_bytree, and gamma. After fine-tuning, we have identified the following optimal values: n_estimators = 352, max_depth = 3, learning_rate = 0.06, colsample_bytree = 0.82, gamma = 0.1. During the model construction process, we employed a rigorous technique known as five-fold cross-validation. This validation method involves randomly dividing the initial training set into five equal parts. Four of these parts are used as the training set, while the remaining part serves as the validation set. This process is repeated five times, with each part serving as the validation set once. This cross-validation approach allows for a more comprehensive evaluation of the model's performance, effectively mitigating issues of overfitting and underfitting. Consequently, it ensures the stability and reliability of the model in practical applications. Through this approach, we were able to continuously optimize and adjust model parameters during the construction process, ultimately arriving at a superior and highly generalizable model.

The model was evaluated based on accuracy, Area Under the Curve (AUC), sensitivity (Sen), and specificity (Spe). After comprehensive assessment of these metrics, the most outstanding model was selected as the final predictive model. To appraise the clinical utility of the model, a test set decision curve was plotted, alongside a calibration curve to assess the model's accuracy. A nomogram was also created to provide a more intuitive visualization of the model's practicality, which was constructed in this experimental context. In Python, the lifelines library is used to plot the decision curve, while the nomogram library is used to plot the Nomogram.

2.5. Immune infiltration landscape

Within the R environment, the "single-sample gene set enrichment analysis (ssGSEA) method" [25] was employed to calculate the abundance of immune infiltrating cells and the activity of immune pathways in the enrolled samples. ssGSEA allows for the

computation of enrichment scores, indicating the absolute level of gene set enrichment in individual samples from a given dataset. Subsequently, the correlation between immune infiltrating cells and immune pathways within the enrolled samples was analyzed. Moreover, the differences in immune infiltration between normal samples and PSS samples were further evaluated. In R, the gsva function from the GSVA (Gene Set Variation Analysis) package is used to perform ssGSEA analysis.

2.6. Statistical analysis

Statistical analyses were conducted using R (4.0.2, http://www.r-project.org) and Python (3.8.5 , https://www.python.org/downloads/release/python-380/). Data normality was checked using Kolmogorov-Smirnov tests. Data that followed a normal distribution are presented as mean ± standard deviation ($\overline{x} \pm s$) and were analyzed using an independent *t*-test. Data that did not adhere to a normal distribution are reported as median (interquartile range) and were assessed using the Mann-Whitney *U* test. The nonparametric Spearman test was utilized to ascertain correlations between variables. LASSO regression was employed for the selection of key genes. To assess the predictive power of the model, various metrics such as accuracy, AUC, sensitivity, and specificity were computed. Additionally, the clinical utility of the model was evaluated through a decision curve analysis.

3. Result

The research methodology is illustrated in Fig. 1. The primary dataset (Dataset I) for this study consisted of 129 samples, including 71 PSS samples and 58 normal samples, all sourced from the GEO database. An external independent validation dataset (Dataset II) included 60 samples, comprising 39 PSS samples and 21 normal samples, which were also obtained entirely from the GEO database. Additionally, Supplementary Table 1 provided a list of 29 m7G-related genes.

3.1. Determination of the m7G-related genes expression matrix

In the present study, we extracted the expression matrix of 29 m7G-related genes from Dataset I, which comprised a total of 129 samples (see Supplementary Table 2). The correlation analysis among these 29 genes is depicted in Fig. 2(A). The results revealed that NCBP3 and EIF3D exhibited the highest positive correlation (R = 0.676), followed closely by a positive correlation of 0.592 between LSM1 and EIF4E. Additionally, NCBP2 and EIF4E1 displayed a positive correlation of 0.546. Conversely, NCBP2 had the highest negative correlation with AGO2, at −0.391. It is evident from our study that no gene pairs exhibited a correlation coefficient greater than 0.9, thus ruling out the presence of collinear redundancy and ensuring that our analysis is not confounded by spurious relationships.

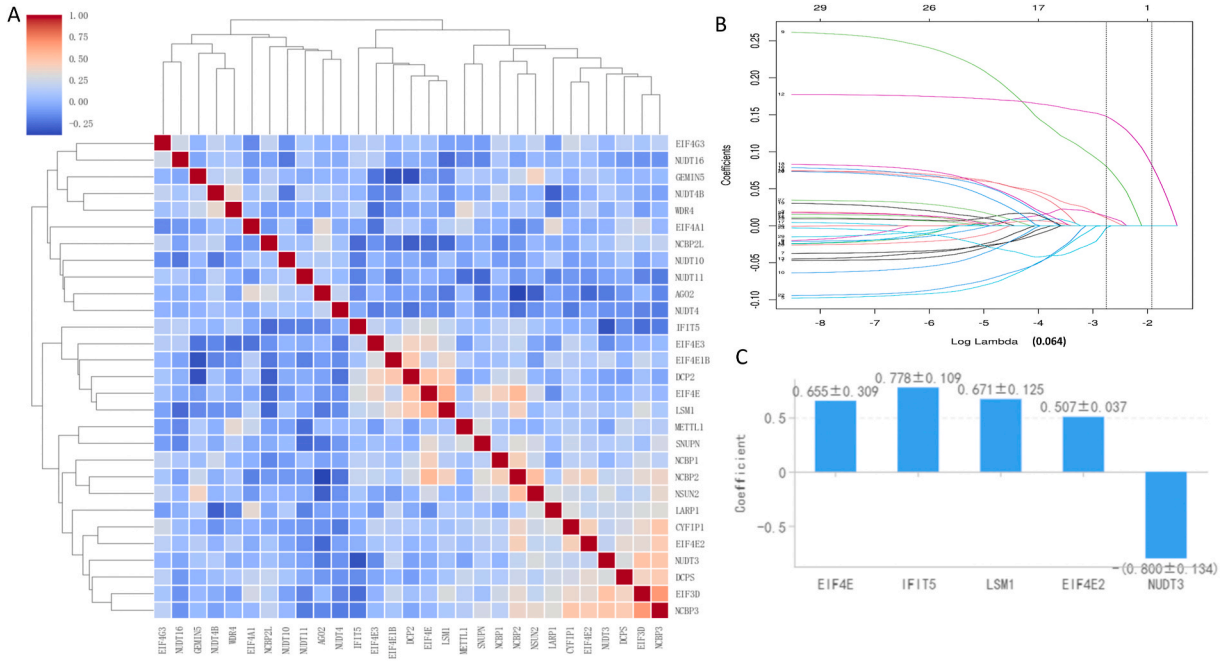


Fig. 2. Key genes selection. (A), Spearman correlation analysis among 29 m7G-realted genes; (B), LASSO coefficient profiles of the selected m7G-realted genes; (C), Distribution of key m7G-realted genes coefficients.

3.2. Screening of the key genes

To assess the concentration levels of genes within the training set, the Spearman correlation analysis was utilized. In cases where the correlation between any two genes surpassed the threshold of 0.9, one gene from the pair was omitted, as illustrated in Fig. 2(A). Specific information can be found in Supplementary Table 2. Following this, LASSO regression analysis was performed using a lambda value of 0.064, as demonstrated in Fig. 2(B). This process facilitated the discovery of five key genes: IFIT5, EIF4E2, LSM1, EIF4E, and NUDT3. The relationship between these five genes and PSS is illustrated in Fig. 2(C). The expression distribution of these five genes between normal and PSS states is presented in Supplementary Fig. 1. Among them, IFIT5, LSM1, EIF4E, and NUDT3 demonstrate significant differential expression. After conducting a significance analysis on the coefficients of these five key genes in the model (as detailed in Supplementary Table 3), we observed that the P-values for the intercept term b and EIF4E were greater than 0.05, indicating that they do not have a statistically significant relationship with the occurrence of PSS. However, the P-values for the remaining four genes were all less than 0.05, strongly suggesting a significant statistical association between these genes and PSS risk prediction. By calculating the mean and standard deviation of the coefficients of these five genes and the intercept term b under various conditions, a risk score for PSS can be represented using the following equation.

$$\begin{aligned}\text{Risk\_Score} = (0.655 \pm 0.309) \times \text{EIF4E} + (0.507 \pm 0.037) \times \text{EIF4E2} + (0.778 \pm 0.109) \times \text{IFIT5} + (0.671 \pm 0.125) \\ \times \text{LSM1-} (0.800 \pm 0.134) \times \text{NUDT3-}(10.028 \pm 0.216)\end{aligned}$$

3.3. Establishment of the prediction model

In this investigation, a predictive model for the incidence of PSS was crafted utilizing five distinct machine learning techniques: SVM, KNN, RF, LightGBM, and XGBoost. Additionally, Dataset II was employed as an independent external dataset to perform external validation of the model. Upon examining Table 1, it is observed that XGBoost outperforms the other four models in terms of accuracy, AUC, and sensitivity on both the test set and the external test set. XGBoost only falls slightly behind KNN in specificity. Upon a closer examination of the calibration curves depicted in Fig. 3(A) and (B), it becomes evident that the XGBoost model exhibits superior performance, adhering more closely to the ground truth. This advantage is particularly noteworthy when considering that the cutoff value for XGBoost was set at 0.624, which likely contributed to its improved accuracy. The overall performance of the LightGBM model lags behind the other four models, indicating that it may be less adept at handling the specific dataset employed in this study. This inferiority could stem from the unique characteristics of the data or the comparison with other models. On the other hand, the RF model exhibits a lower Accuracy compared to KNN and XGBoost, suggesting it may not be the most effective choice for this particular evaluation metric. Furthermore, the SVM model's Sensitivity on the Extra-test set stands at 0.636, significantly below the 0.783 achieved by XGBoost, indicating a weaker ability to detect positive cases accurately. Furthermore, the DCA curve corroborates the clinical utility of the XGBoost model, as seen in Fig. 3 (C) . When the threshold (T) ranges from 30 % to 100 %, the baseline model's net benefit ranges from 0.0 to 0.35. This indicates that when using this model in a clinical setting, for every 100 patients, there will be a benefit for 0 to 35 individuals. The XGBoost model achieved an accuracy of 0.831(0.783–0.879), AUC of 0.922(0.873–0.971), sensitivity of 0.834(0.789–0.880), and specificity of 0.875(0.825–0.924) in the training set with a cutoff of 0.624(0.590–0.658). Within the test set, these metrics were also recorded as 0.781(0.764–0.798), 0.889(0.828–0.950), 0.838(0.754–0.921), and 0.776 (0.697–0.855), respectively. In the external test set, the performance indicators were 0.723 for accuracy, 0.791 for AUC, 0.783 for sensitivity, and 0.811 for specificity. Finally, through the nomograms (Fig. 3(D)), we can easily assess the PSS risk probability of a suspicious individual.

Table 1
The performance of five different prediction models.

Model	Task	Accuracy (95%CI)	AUC (95%CI)	Cutoff (95%CI)	Sensitivity (95%CI)	Specificity (95%CI)
SVM	Train	0.783 (0.770–0.796)	0.871 (0.805–0.938)	0.592 (0.533–0.651)	0.732 (0.686–0.779)	0.875 (0.808–0.942)
	Test	0.706 (0.647–0.765)	0.810 (0.638–0.979)	0.592 (0.533–0.651)	0.789 (0.689–0.888)	0.794 (0.629–0.959)
	Extra-test	0.700	0.717	0.592	0.636	0.778
KNN	Train	0.670 (0.614–0.727)	0.835 (0.762–0.909)	0.667 (0.563–0.770)	0.750 (0.680–0.819)	0.741 (0.691–0.791)
	Test	0.660 (0.589–0.730)	0.763 (0.585–0.939)	0.667 (0.563–0.770)	0.593 (0.420–0.767)	0.797 (0.607–0.987)
	Extra-test	0.560	0.742	0.667	0.617	0.838
RF	Train	0.777 (0.709–0.846)	0.898 (0.835–0.961)	0.621 (0.486–0.761)	0.767 (0.678–0.856)	0.878 (0.816–0.940)
	Test	0.696 (0.585–0.806)	0.814 (0.632–0.984)	0.621 (0.486–0.761)	0.800 (0.702–0.898)	0.791 (0.668–0.915)
	Extra-test	0.641	0.738	0.621	0.650	0.684
XGBoost	Train	0.831 (0.783–0.879)	0.922 (0.873–0.971)	0.624 (0.590–0.658)	0.834 (0.789–0.880)	0.875 (0.825–0.924)
	Test	0.781 (0.764–0.798)	0.889 (0.828–0.950)	0.624 (0.590–0.658)	0.838 (0.754–0.921)	0.776 (0.697–0.855)
	Extra-test	0.723	0.791	0.624	0.783	0.811
LightGBM	Train	0.566 (0.433–0.698)	0.713 (0.621–0.805)	0.718 (0.350–0.986)	0.623 (0.544–0.703)	0.841 (0.781–0.894)
	Test	0.552 (0.405–0.701)	0.671 (0.475–0.868)	0.718 (0.350–0.986)	0.561 (0.414–0.708)	0.864 (0.779–0.948)
	Extra-test	0.641	0.536	0.718	0.538	0.758

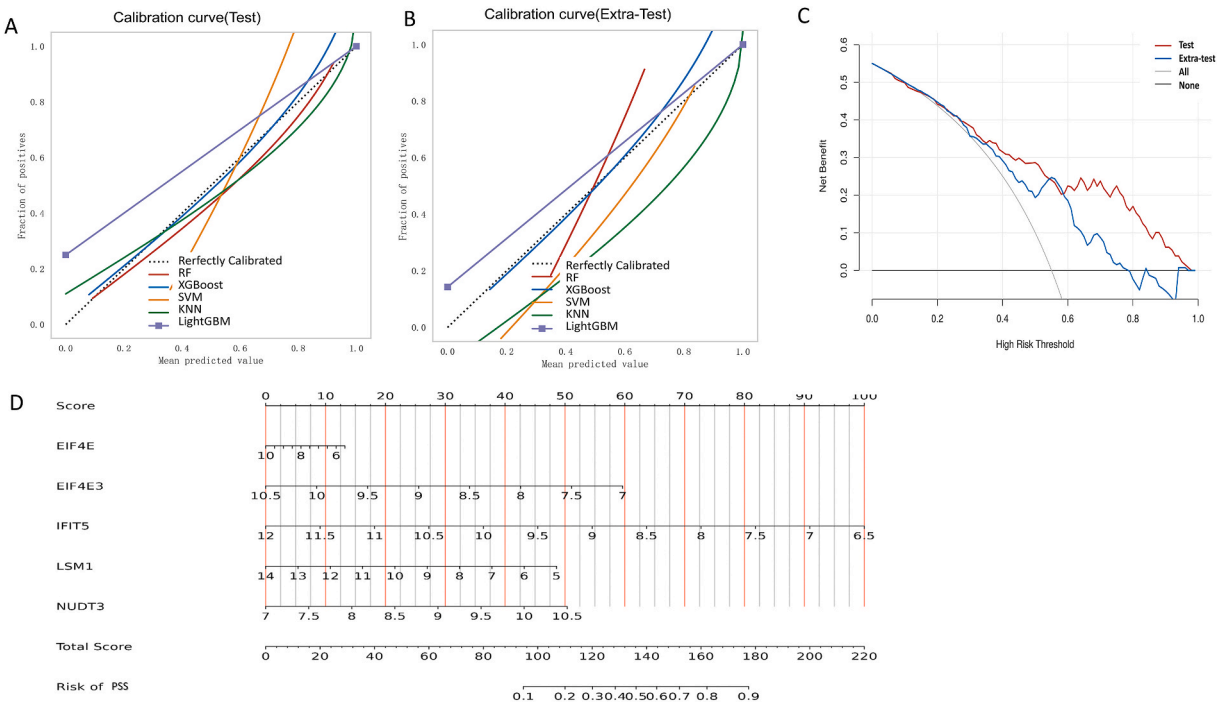


Fig. 3. The calibration curves of the five models in the training group (A) and the test group (B); (C) Decision curve analysis of the XGBoost in the test group (red line) and extra-test group (blue line); (D) Nomograms for predicting the risk of primary sjögren's syndrome. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

3.4. Immune infiltration results

The ssGSEA method was used to evaluate the abundance of immune infiltration cells in PSS samples. Fig. 4 indicated that $CD8^+$ T cells, neutrophils, and tumor infiltrating lymphocytes (TIL) exhibited high abundance in the enrolled samples. Additionally, the included samples demonstrated high activity of cytolytic activity, human leukocyte antigens (HLA), major histocompatibility complex (MHC) class I, and Type I interferon response. Fig. 4(B) indicated a positive correlation between immune infiltration cells and immune function. TIL exhibited a strong positive correlation with the other immune infiltration cells and immune function, with the strongest correlation observed between TIL and B cells (R = 0.66, Fig. 4(B)). The correlation analysis of immune pathways (Fig. 4(B)) yielded

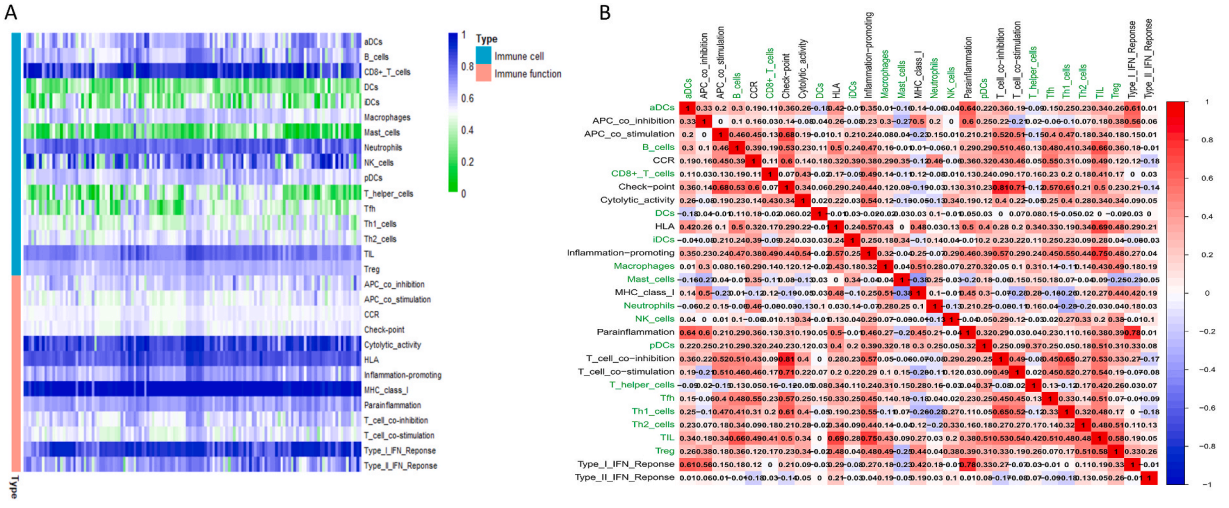


Fig. 4. (A) Heatmap showing immune infiltration cells and immune function among PSS; (B) Immune infiltration cell (green label) and immune function(black label) correlation analysis. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

similar results to that of immune infiltration cells. Checkpoint and T cell co-inhibition demonstrated the strongest positive correlation (R = 0.81). Activated dendritic cells (aDCs), B cells, and mast cells were the only immune infiltration cells that exhibited differences between the experimental and control groups (Fig. 4(B)). Seven out of the 13 immune pathways exhibited differences between the normal group and the PSS group, with these immune function being more active in the PSS group (Fig. 5(A)). Fig. 5(B) suggests that four of the five key genes (IFIT5, EIF4E2, LSM1, and EIF4E) utilized for modeling demonstrate significant associations with immune infiltration. However, it is worth noting that the strength of these associations may vary, and further analysis is required to fully understand their individual contributions and relationships with immune infiltration.

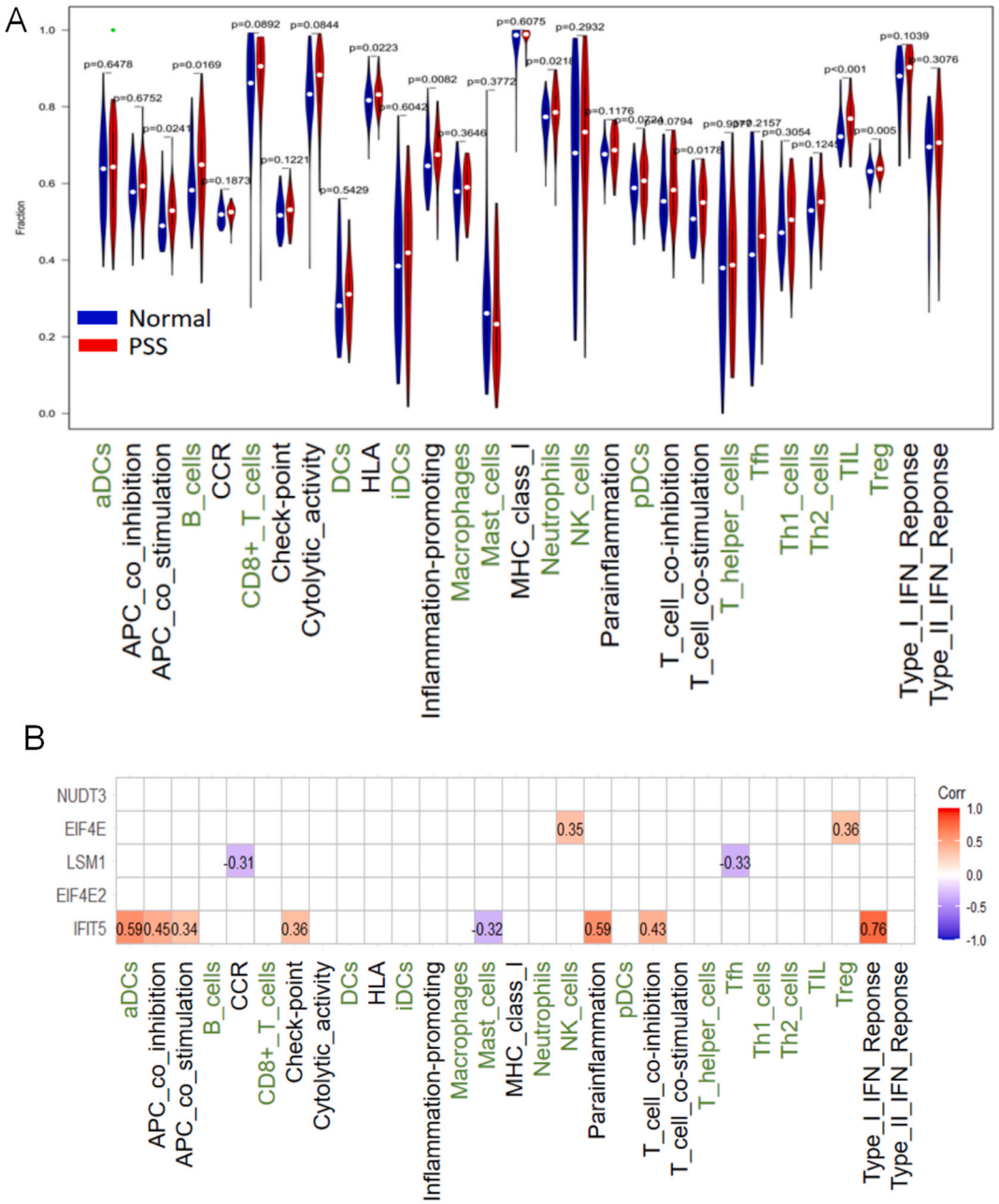


Fig. 5. (A) Difference analysis of immune infiltration cell (green label) and immune function(black label) between PSS and normal groups; (B) Correlation analysis betwwen five key m7G-realted genes and immune infiltration. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

4. Discussion

PSS, a common manifestation in autoimmune diseases, can affect multiple organs and poses a threat to patients' lives [1]. Current management lacks specificity, especially in organ involvement, leading to reduced treatment efficacy. Therefore, early detection and timely treatment are crucial. Studies have shown the etiology of PSS, explained the molecular mechanism behind its pathogenesis, and emphasized the potential for diagnosing and treating PSS with high-throughput sequencing technology advancements [26].

Machine learning has been shown to achieve superior results in processing multidimensional data [27]. Moreover, the combination of machine learning methods in genomics has proven to be more effective than traditional statistical methods [28]. The process of m7G methylation plays a crucial role in regulating various mechanisms of human life [11]. For this study, we focused on the expression of m7G-related genes in both the normal and PSS groups. Initially, we screened five key genes - IFIT5, EIF4E2, LSM1and EIF4E, and NUDT3 - by employing mathematical methods. Utilizing a binary logistic regression model to assess the coefficients' significance of these five genes in relation to PSS revealed that all except EIF4E exhibited significant and favorable associations, thereby validating the correctness of our choice of core genes in this study. By considering the impact of various conditions on the coefficients of the core genes in LASSO regression, we have derived a PSS risk scoring equation that possesses high credibility, robustness, and generalization capabilities. Subsequently, we utilized five machine learning algorithms, namely SVM, KNN, RF, LightGBM, and XGBoost, to construct a risk prediction model for PSS. The models were thoroughly evaluated using a range of metrics, including accuracy, AUC, sensitivity, and specificity. In this study, the choice of these metrics was informed by their extensive application in binary classification tasks and their capacity to offer diverse perspectives on model performance. Accuracy offers a measure of the model's ability to correctly classify samples, while AUC quantifies its performance across various classification thresholds. Additionally, sensitivity and specificity reflect the model's proficiency in identifying positive and negative samples, respectively, which is particularly crucial in PSS prediction, as our goal is to accurately detect PSS patients without misclassifying healthy individuals. Given their comprehensive assessment of the model's classification capabilities, we believe that these metrics are highly suitable for PSS prediction, providing valuable insights for practical applications. After a rigorous evaluation, we selected the optimal model to predict the risk of PSS, ensuring a reliable and accurate tool for clinical decision-making.

In this study, after thorough comparison and evaluation, the XGBoost model has demonstrated superior performance across various performance metrics, significantly outperforming four other machine learning models including SVM, KNN, RF, and LightGBM. Specifically, the XGBoost model not only led in key evaluation criteria such as accuracy, AUC, sensitivity and specificity but also showed remarkable advantages in terms of model stability and generalization capabilities. These results indicate that the XGBoost model, with its high degree of flexibility and powerful performance, possesses greater efficiency and accuracy when dealing with the datasets in this study. In contrast, although SVM, KNN, RF, and LightGBM are widely used machine learning algorithms with commendable performances in their respective application scenarios, they did not achieve the optimal efficacy level as XGBoost under the specific conditions of this experiment. SVM, for instance, is renowned for its ability to handle high-dimensional data and its robustness to outliers, but it may suffer from sensitivity to kernel function selection and parameter tuning [29]. KNN, on the other hand, is a simple yet effective method for classification and regression, but its performance can be heavily influenced by the choice of the number of neighbors and the distance metric [30]. RF, as an ensemble method, often achieves good performance by combining the predictions of multiple decision trees. However, its performance can vary depending on the number and depth of the trees, as well as the sampling strategy used [31]. Similarly, LightGBM, a gradient boosting framework optimized for speed and efficiency, is a strong contender in many machine learning competitions. However, its performance may be limited by the choice of hyperparameters and the nature of the data [32]. The success of the XGBoost model in this experiment can be attributed to its unique gradient boosting mechanism and effective control over model complexity. XGBoost's ability to prevent overfitting while capturing subtle patterns in the data, along with its optimized algorithms for handling large-scale datasets [33], has provided a guarantee for its efficient computation and superior performance compared to the other algorithms mentioned. The findings of this study highlight the importance of considering data characteristics and task requirements when selecting a machine learning model. XGBoost, with its good performance, stands out in this research and offers a strong reference for future studies on similar issues. However, we also recognize that no single model can be optimal in all situations; therefore, in practical applications, it is still necessary to choose and adjust the model according to the specific context.

LASSO regression is a penalized estimator for collinear data, achieving variable selection by zeroing some coefficients [34]. XGBoost, a gradient boosting algorithm based on decision trees, not only demonstrates significant advantages in handling large-scale datasets [33] but also remains effective for small to medium-sized datasets. It achieves a balance between performance and speed, making it a popular choice in data science [35,36]. The XGBoost model demonstrated excellent predictive performance across the training set, the test set, and the external test set (Table 1). Observing the 95 % CI of various indicators from the test set, the model established through five-fold cross-validation has demonstrated good robustness. This result indicates that the model maintains relatively stable performance across different data partitions, providing a reliable basis for practical applications. DCA indicated the model's clinical utility (Fig. 3(C)).

In the medical field, it is not only necessary for the model to achieve high prediction accuracy on the test set, but also important to identify the significant factors influencing the onset and progression of the disease. Our study revealed a close relationship between the m7G regulatory factor and various immune infiltration cells and pathways. The study demonstrated higher immunological activity in PSS patients for aDCs, B cells, Mast cells, APC co-stimulation, Check-point, Parainflammation, T-cell-co-inhibition, and Type-I–IFN–Reponse (Fig. 5(A)). Numerous studies have demonstrated the significant involvement of immune cells and related cytokines from both the innate and adaptive immune systems in the onset of PSS [37]. Currently, it is widely accepted that environmental factors, including viruses, can initiate a cascade of reactions that lead to inflammation and the subsequent involvement of exocrine glands and

systems [38]. The Type I IFN pathway, Toll-like receptors (TLRs), IL-1 family cytokines, Dendritic cells (DCs), and NK cells all have significant roles in innate immunity [39]. The study revealed the hyperactivity of the type I IFN pathway in PSS patients (Fig. 4(A)), and its strong correlation with inflammatory response (Fig. 4(B)), confirming previous findings [40,41]. It is known that the main characteristic of PSS is the infiltration of a large number of lymphocytes, primarily $CD4^+$ T cells, in the exocrine glands [42]. Furthermore, the study observed significantly higher expression of B cells and aDCs in PSS patients compared to normal individuals (Fig. 5(A)). For instance, the study identified correlations between IFIT5 and aDCs ($R = 0.59$), as well as IFIT5 and Type I IFN Response ($R = 0.76$) (Fig. 5(B)). Consequently, the study suggests that IFIT5, EIF4E2, LSM1and EIF4E, and NUDT3 identified in this investigation can serve as biomarkers for the onset and prognosis of PSS.

Consistent with our findings, research has confirmed that IFIT5 is significantly elevated in autoimmune diseases [43,44]. The IFIT5 gene encodes a protein that plays a crucial role in the interferon signaling pathway. This protein is primarily expressed in immune cells and regulates cellular processes such as growth, differentiation, and apoptosis.Research findings suggest that IFIT5 may be associated with the autoimmune response involved in PSS. A study has reported that IFIT5, as a member of the interferon-induced tetratricopeptide repeat family, enhances the phosphorylation of i-κB kinase (IKK) and activation of NF-κB by interacting with Transforming Growth Factor β-Activated Kinase 1 (TAK1) and IKK. This underscores the role of IFIT5 in innate immunity [45]. In the case of IFIT5, when the host is infected by a virus, it induces an antiviral response in host cells [46]. The expression of MHC class I is primarily regulated by NF-κB and IFN-regulatory factors in response to IFN-γstimulation [47]. In our current study, we have also uncovered the aberrant upregulation of MHC class I in PSS. Evidently, IFIT5 plays a significant role in immune regulation within the human body. NUDT3, a Nudix protein, has been proposed to possess mRNA cleavage activity in cells and acts as a regulator of migration in MCF-7 breast cancer cells [48]. However, its role in autoimmune diseases remains unclear. Our research has found that it is significantly down-regulated in PSS, acting as a protective factor against the onset of PSS, yet its association with the immune system is not particularly strong. EIF4E has been demonstrated to participate in the translation of viral mRNA and the proliferation of infected cells [49]. SARS-CoV-2 can effectively replicate in target cells by utilizing the ERK/MNK1/EIF4E pathway [50,51]. Extensive research has found that EIF4E2 plays a crucial role in tumor progression and metastasis [52]. Furthermore, Yang et al. demonstrated that high EIF4E2 expression serves as an independent prognostic risk factor for patients with uveal melanoma (UM) [53]. Throughout the course of UM, EIF4E2 may play a pivotal role in hypoxia-related signaling pathways. Additional research has uncovered a regulatory mechanism by which EIF4E2 acts as a potential tumor suppressor, inhibiting HIF-2- and EIF4E2-mediated translation activation of oncogenic mRNAs. Notably, DEAD Box Protein Family Member (DDX28) has been identified as a negative regulator of hypoxia-inducible factor 2α- and eukaryotic initiation factor 4E2-directed hypoxic translation [53]. Additionally, Melanson Gaelan et al. discovered the EIF4E2-directed hypoxic cap-dependent translation machinery, revealing novel therapeutic potential for cancer treatment [54]. The expression of LSM1 in pancreatic cancer cells leads to accelerated growth, decreased sensitivity to chemotherapy drugs, and enhanced migration and invasion capabilitie [55]. This upregulation affects key genes regulating apoptosis, metastasis, and epithelial-mesenchymal transition (EMT), which is consistent with LSM1's function in mRNA stability and processing [56]. In the context of autoimmune diseases, EMT plays a crucial role. For instance, in rheumatoid arthritis (RA), synovial cells undergo EMT, acquiring mesenchymal characteristics and invading the joint cavity, which further promotes inflammation and joint destruction [57]. Similarly, in systemic lupus erythematosus (SLE), the epithelial cells of the glomerulus undergo EMT, leading to kidney tissue damage and the appearance of proteinuria [58]. Therefore, LSM1 may be involved in the pathological processes of autoimmune diseases such as rheumatoid arthritis, systemic lupus erythematosus, and PSS by affecting the expression of genes related to EMT.

In conclusion, the predictive model developed from m7G-related genes offers promising clinical applications. The involvement of m7G methylation in PSS pathogenesis suggests its role in autoimmune and adaptive immune responses, contributing to disease progression. However, this study has limitations, such as relying solely on public datasets without experimental validation and a relatively small sample size, which could lead to overfitting. Furthermore, in our feature engineering process, we primarily relied on purely mathematical methods to select core genes, neglecting the potential interactions between genes and their specific biological functions in the context of PSS occurrence. These concerns should be addressed in future studies.

5. Conclusion

In this research, we harnessed m7G-related genes to develop a risk prediction model for the onset of PSS utilizing machine learning methodologies. We then assessed the efficacy of this model and found that the XGBoost algorithm excelled in predicting PSS, indicating its potential utility in clinical settings. Moreover, our results implied that m7G methylation might be involved in the development and progression of PSS through immune response pathways.

Funding

This study was supported by.

1Key Laboratory of Tumor Precision Medicine, Hunan colleges and Universities Project (2019-379)
2A Project Supported by Scientific Research Fund of Hunan Provincial Education Department(20k118)
3.Macao Polytechnic University, Grant number: RP/FCA-15/2022.

Availability of data and materials

Not applicable.

Conflict of interest statement and consent for publication

The authors have no ethical, legal and financial conflicts related to the article. All authors read and approved the manuscript to publish.

CRediT authorship contribution statement

Hui Xie: Writing – review & editing. **Yin-mei Deng:** Writing – original draft, Data curation. **Jiao-yan Li:** Writing – original draft. **Kai-hong Xie:** Writing – original draft. **Tan Tao:** Writing – review & editing. **Jian-fang Zhang:** Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Not applicable.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2024.e31307.

References

[1] C.Y. Hsu, K.C. Hung, M.S. Lin, C.H. Ko, Y.S. Lin, T.H. Chen, C.Y. Lin, Y.C. Chen, The effect of pilocarpine on dental caries in patients with primary Sjögren's syndrome: a database prospective cohort study, Arthritis Res. Ther. 21 (1) (2019 Nov 27) 251, https://doi.org/10.1186/s13075-019-2031-7. PMID: 31775834; PMCID: PMC6882320.
[2] P. Brito-Zerón, C. Baldini, H. Bootsma, S.J. Bowman, R. Jonsson, X. Mariette, K. Sivils, E. Theander, A. Tzioufas, M. Ramos-Casals, Sjögren syndrome, Nat. Rev. Dis. Prim. 2 (2016 Jul 7) 16047, https://doi.org/10.1038/nrdp.2016.47. PMID: 27383445.
[3] O. Aiyegbusi, L. McGregor, L. McGeoch, et al., Renal disease in primary sjögren's syndrome, Rheumatol Ther 8 (1) (2021) 63–80.
[4] C.P. Mavragani, H.M. Moutsopoulos, The geoepidemiology of sjögren's syndrome, Autoimmun. Rev. 9 (5) (2010) A305–A310.
[5] N. Li, Y.S. Li, J.W. Hu, et al., A link between mitochondrial dysfunction and the immune microenvironment of salivary glands in primary sjogren's syndrome, Front. Immunol. 13 (2022) 845209.
[6] J. Imgenberg-Kreuz, A. Rasmussen, K. Sivils, G. Nordmark, Genetics and epigenetics in primary Sjögren's syndrome, Rheumatology 60 (2021) 2085–2098, https://doi.org/10.1093/rheumatology/key330.
[7] H.Y. Sun, A.K. Lv, H. Yao, Relationship of miRNA-146a to primary Sjögren's syndrome and to systemic lupus erythematosus: a meta-analysis, Rheumatol. Int. 37 (2017) 1311–1316, https://doi.org/10.1007/s00296-017-3756-8.
[8] P. Boccaletto, M.A. Machnicka, E. Purta, P. Piątkowski, B. Bagiński, T.K. Wirecki, et al., Modomics: a database of rna modification pathways. 2017 Update, Nucleic Acids Res. 46 (2018) D303–D307, https://doi.org/10.1093/nar/gkx1030.
[9] X. Han, M. Wang, Y.L. Zhao, Y. Yang, Y.G. Yang, Rna methylations in human cancers, Semin. Cancer Biol. 75 (2021) 97–115, https://doi.org/10.1016/j.semcancer.2020.11.007.
[10] L.Y. Zhao, J. Song, Y. Liu, C.X. Song, C. Yi, Mapping the epigenetic modifications of DNA and Rna, Protein Cell 11 (2020) 792–808, https://doi.org/10.1007/s13238-020-00733-7.
[11] L. Malbec, T. Zhang, Y.S. Chen, Y. Zhang, B.F. Sun, B.Y. Shi, et al., Dynamic methylome of internal Mrna N(7)-Methylguanosine and its regulatory role in translation, Cell Res. 29 (2019) 927–941, https://doi.org/10.1038/s41422-019-0230-z.
[12] B. Chen, W. Jiang, Y. Huang, J. Zhang, P. Yu, L. Wu, H. Peng, N(7)-methylguanosine tRNA modification promotes tumorigenesis and chemoresistance through WNT/β-catenin pathway in nasopharyngeal carcinoma, Oncogene (2022), https://doi.org/10.1038/s41388-022-02250-9.
[13] J. Chen, K. Li, J. Chen, X. Wang, R. Ling, M. Cheng, Z. Chen, F. Chen, Q. He, S. Li, et al., Aberrant translation regulated by METTL1/WDR4-mediated tRNA N7-methylguanosine modification drives head and neck squamous cell carcinoma progression, Cancer Commun. 42 (3) (2022) 223–244, https://doi.org/10.1002/cac2.12273.
[14] Y. Liu, J. Zhu, L. Ding, Involvement of RNA methylation modification patterns mediated by m7G, m6A, m5C and m1A regulators in immune microenvironment regulation of Sjögren's syndrome, Cell. Signal. 106 (2023 Jun) 110650, https://doi.org/10.1016/j.cellsig.2023.110650. Epub 2023 Mar 17. PMID: 36935085.
[15] Jeffrey T. Leek, W. Evan Johnson, Hilary S. Parker, Andrew E. Jaffe, John D. Storey, The sva package for removing batch effects and other unwanted variation in high-throughput experiments, Bioinformatics 28 (6) (March 2012) 882–883, https://doi.org/10.1093/bioinformatics/bts034.
[16] F. Pedregosa, G. Varoquaux, A. Gramfort, et al., Scikit-learn: machine learning in python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[17] Y. Hirano, K. Ihara, T. Masuda, T. Yamamoto, I. Iwata, A. Takahashi, H. Awata, N. Nakamura, M. Takakura, Y. Suzuki, J. Horiuchi, H. Okuno, M. Saitoe, Shifting transcriptional machinery is required for long-term memory maintenance and modification in Drosophila mushroom bodies, Nat. Commun. 7 (2016 Nov 14) 13471, https://doi.org/10.1038/ncomms13471. PMID: 27841260; PMCID: PMC5114576.
[18] S. Binney, M.K. Person, R.M. Traxler, R. Cook, W.A. Bower, K. Hendricks, Algorithms for the identification of anthrax meningitis during a mass casualty event based on a systematic review of systemic anthrax from 1880 through 2018, Clin. Infect. Dis. 75 (Suppl 3) (2022 Oct 17) S468–S477, https://doi.org/10.1093/cid/ciac546. PMID: 36251554; PMCID: PMC9649431.
[19] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al., Scikit-learn: machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.

[20] P. Bonaventura, V. Alcazer, V. Mutez, L. Tonon, J. Martin, N. Chuvin, E. Michel, R.E. Boulos, Y. Estornes, J. Valladeau-Guilemond, A. Viari, Q. Wang, C. Caux, S. Depil, Identification of shared tumor epitopes from endogenous retroviruses inducing high-avidity cytotoxic T cells for cancer immunotherapy, Sci. Adv. 8 (4) (2022 Jan 28) eabj3671, https://doi.org/10.1126/sciadv.abj3671. Epub 2022 Jan 26. PMID: 35080970; PMCID: PMC8791462.

[21] C. Lee, S. Lee, E. Park, J. Hong, D.Y. Shin, J.M. Byun, H. Yun, Y. Koh, S.S. Yoon, Transcriptional signatures of the BCL2 family for individualized acute myeloid leukaemia treatment, Genome Med. 14 (1) (2022 Sep 28) 111, https://doi.org/10.1186/s13073-022-01115-w. PMID: 36171613; PMCID: PMC9520894.

[22] J.T. Leek, W.E. Johnson, H.S. Parker, A.E. Jaffe, J.D. Storey, The sva package for removing batch effects and other unwanted variation in high-throughput experiments, Bioinformatics 28 (6) (2012 Mar 15) 882–883, https://doi.org/10.1093/bioinformatics/bts034. Epub 2012 Jan 17. PMID: 22257669; PMCID: PMC3307112.

[23] E.A. Orellana, Q. Liu, E. Yankova, et al., METTL1-mediated m7G modification of Arg-TCT tRNA drives oncogenic transformation, Mol. Cell 81 (16) (2021) 3323–3338.e14, https://doi.org/10.1016/j.molcel.2021.06.031.

[24] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, et al., Scikit-learn: machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830, https://doi.org/10.1016/j.patcog.2011.04.006.

[25] Q. Meng, LightGBM: A Highly Efficient Gradient Boosting Decision Tree[C]//Neural Information Processing Systems, Curran Associates Inc., 2017.

[26] T. Chen, C. Guestrin, XGBoost: A Scalable Tree Boosting System, ACM, 2016, https://doi.org/10.1145/2939672.2939785.

[27] D.A. Barbie, P. Tamayo, J.S. Boehm, S.Y. Kim, S.E. Moody, I.F. Dunn, A.C. Schinzel, P. Sandy, E. Meylan, C. Scholl, S. Fröhling, E.M. Chan, M.L. Sos, K. Michel, C. Mermel, S.J. Silver, B.A. Weir, J.H. Reiling, Q. Sheng, P.B. Gupta, R.C. Wadlow, H. Le, S. Hoersch, B.S. Wittner, S. Ramaswamy, D.M. Livingston, D. M. Sabatini, M. Meyerson, R.K. Thomas, E.S. Lander, J.P. Mesirov, D.E. Root, D.G. Gilliland, T. Jacks, W.C. Hahn, Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1, Nature 462 (7269) (2009 Nov 5) 108–112, https://doi.org/10.1038/nature08460. Epub 2009 Oct 21. PMID: 19847166; PMCID: PMC2783335.

[28] Y. Zhang, P. Batys, J.T. O'Neal, F. Li, M. Sammalkorpi, J.L. Lutkenhaus, Molecular origin of the glass transition in polyelectrolyte assemblies, ACS Cent. Sci. 4 (5) (2018 May 23) 638–644, https://doi.org/10.1021/acscentsci.8b00137. Epub 2018 Apr 13. PMID: 29806011; PMCID: PMC5968513.

[29] A. Roy, S. Chakraborty, Support vector machine in structural reliability analysis: a review, Reliab. Eng. Syst. Saf. 233 (2023) 109126.

[30] U.G. Inyang, F.F. Ijebu, F.B. Osang, A.A. Afolorunso, S.S. Udoh, I.J. Eyoh, A dataset-driven parameter tuning approach for enhanced K-nearest neighbour algorithm performance, Int. J. Adv. Sci. Eng. Inf. Technol. 13 (1) (2023).

[31] R. Banik, A. Biswas, Improving solar PV prediction performance with RF-CatBoost ensemble: a robust and complementary approach, Renewable Energy Focus 46 (2023) 207–221.

[32] U. Allende, Application of Shallow Neural Networks to Retail Intermittent Demand Time Series, 2023.

[33] L. Zhang, J. Zhang, Q. Nie, DIRECT-NET: an efficient method to discover cis-regulatory elements and construct regulatory networks from single-cell multiomics data, Sci. Adv. 8 (22) (2022 Jun 3) eabl7393, https://doi.org/10.1126/sciadv.abl7393. Epub 2022 Jun 1. PMID: 35648859; PMCID: PMC9159696.

[34] C. Jiang, Y. Fu, G. Liu, B. Shu, J. Davis, G.K. Tofaris, Multiplexed profiling of extracellular vesicles for biomarker development, Nano-Micro Lett. 14 (1) (2021 Dec 2) 3, https://doi.org/10.1007/s40820-021-00753-w. PMID: 34855021; PMCID: PMC8638654.

[35] R.M. Hoogeveen, J.P.B. Pereira, N.S. Nurmohamed, V. Zampoleri, M.J. Bom, A. Baragetti, S.M. Boekholdt, P. Knaapen, K.T. Khaw, N.J. Wareham, A.K. Groen, A. L. Catapano, W. Koenig, E. Levin, E.S.G. Stroes, Improved cardiovascular risk prediction using targeted plasma proteomics in primary prevention, Eur. Heart J. 41 (41) (2020 Nov 1) 3998–4007, https://doi.org/10.1093/eurheartj/ehaa648. PMID: 32808014; PMCID: PMC7672529.

[36] N.P. Stone, G. Demo, E. Agnello, B.A. Kelch, Principles for enhancing virus capsid capacity and stability from a thermophilic virus capsid structure, Nat. Commun. 10 (1) (2019 Oct 2) 4471, https://doi.org/10.1038/s41467-019-12341-z. PMID: 31578335; PMCID: PMC6775164.

[37] E. Shrock, E. Fujimura, T. Kula, R.T. Timms, I.H. Lee, Y. Leng, M.L. Robinson, B.M. Sie, M.Z. Li, Y. Chen, J. Logue, A. Zuiani, D. McCulloch, F.J.N. Lelis, S. Henson, D.R. Monaco, M. Travers, S. Habibi, W.A. Clarke, P. Caturegli, O. Laeyendecker, A. Piechocka-Trocha, J.Z. Li, A. Khatri, H.Y. Chu, MGH COVID-19 Collection & Processing Team, Viral epitope profiling of COVID-19 patients reveals cross-reactivity and correlates of severity, in: A.C. Villani, K. Kays, M. B. Goldberg, N. Hacohen, M.R. Filbin, X.G. Yu, B.D. Walker, D.R. Wesemann, H.B. Larman, J.A. Lederer, S.J. Elledge (Eds.), Science 370 (6520) (2020 Nov 27) eabd4250, https://doi.org/10.1126/science.abd4250. Epub 2020 Sep 29. PMID: 32994364; PMCID: PMC7857405.

[38] S. Liu, R.Y. Patel, P.R. Daga, H. Liu, G. Fu, R.J. Doerksen, Y. Chen, D.E. Wilkins, Combined rule extraction and feature elimination in supervised classification, IEEE Trans. NanoBioscience 11 (3) (2012 Sep) 228–236, https://doi.org/10.1109/TNB.2012.2213264. PMID: 22987128; PMCID: PMC6295448.

[39] S. Kivity, M.T. Arango, M. Ehrenfeld, et al., Infection and autoimmunity in Sjogren's syndrome: a clinical study and comprehensive review, J. Autoimmun. 51 (2014) 17–22, https://doi.org/10.1016/j.jaut.2014.02.008.

[40] S. Colafrancesco, C. Ciccacci, R. Priori, A. Latini, G. Picarelli, F. Arienzo, G. Novelli, G. Valesini, C. Perricone, P. Borgiani, STAT4, TRAF3IP2, IL10, and HCP5 polymorphisms in sjögren's syndrome: association with disease susceptibility and clinical aspects, J Immunol Res 2019 (2019 Feb 10) 7682827, https://doi.org/10.1155/2019/7682827. PMID: 30882006; PMCID: PMC6387711.

[41] A. Nezos, C.P. Mavragani, Contribution of genetic factors to Sjögren's syndrome and Sjögren's syndrome related lymphomagenesis, Journal of Immunology Research 2015 (2015) 12, https://doi.org/10.1155/2015/754825.754825。.

[42] X. Lin, X. Wang, F. Xiao, K. Ma, L. Liu, X. Wang, D. Xu, F. Wang, X. Shi, D. Liu, Y. Zhao, L. Lu, IL-10-producing regulatory B cells restrain the T follicular helper cell response in primary Sjögren's syndrome, Cell. Mol. Immunol. 16 (12) (2019 Dec) 921–931, https://doi.org/10.1038/s41423-019-0227-z. Epub 2019 Apr 4. PMID: 30948793; PMCID: PMC6884445.

[43] J. Jia, H. Shi, M. Liu, T. Liu, J. Gu, L. Wan, J. Teng, H. Liu, X. Cheng, J. Ye, Y. Su, Y. Sun, W. Gong, C. Yang, Q. Hu, Cytomegalovirus infection may trigger adult-onset still's disease onset or relapses, Front. Immunol. 10 (2019 Apr 24) 898, https://doi.org/10.3389/fimmu.2019.00898. PMID: 31068953; PMCID: PMC6491741.

[44] L. Zhang, P. Xu, X. Wang, Z. Zhang, W. Zhao, Z. Li, G. Yang, P. Liu, Identification of differentially expressed genes in primary Sjögren's syndrome, J. Cell. Biochem. 120 (10) (2019 Oct) 17368–17377, https://doi.org/10.1002/jcb.29001. Epub 2019 May 24. PMID: 31125139.

[45] C. Zheng, Z. Zheng, Z. Zhang, J. Meng, Y. Liu, X. Ke, IFIT5 positively regulates NF-kappaB signaling through synergizing the recruitment of IkappaB kinase (IKK) to TGF-beta-activated kinase 1 (TAK1), Cell. Signal. 27 (2015) 2343–2354.

[46] B. Zhang, X. Liu, W. Chen, L. Chen, IFIT5 potentiates anti-viral response through enhancing innate immune signaling pathways, Acta Biochim. Biophys. Sin. 45 (10) (2013) 867–874, https://doi.org/10.1093/abbs/gmt088.

[47] S.K. Garg, E.A. Welsh, B. Fang, Y.I. Hernandez, T. Rose, J. Gray, J.M. Koomen, A. Berglund, J.J. Mulé, J. Markowitz, Multi-omics and informatics analysis of FFPE tissues derived from melanoma patients with long/short responses to anti-PD1 therapy reveals pathways of response, Cancers 12 (12) (2020 Nov 26) 3515, https://doi.org/10.3390/cancers12123515. PMID: 33255891; PMCID: PMC7768436.

[48] E. Grudzien-Nogalska, X. Jiao, M.G. Song, R.P. Hart, M. Kiledjian, Nudt3 is an mRNA decapping enzyme that modulates cell migration, RNA 22 (5) (2016 May) 773–781, https://doi.org/10.1261/rna.055699.115. Epub 2016 Mar 1. PMID: 26932476; PMCID: PMC4836651.

[49] W.R. Strohl, Z. Ku, Z. An, S.F. Carroll, B.A. Keyt, L.M. Strohl, Passive immunotherapy against SARS-CoV-2: from plasma-based therapy to single potent antibodies in the race to stay ahead of the variants, BioDrugs 36 (3) (2022 May) 231–323, https://doi.org/10.1007/s40259-022-00529-7. Epub 2022 Apr 27. PMID: 35476216; PMCID: PMC9043892.

[50] P. Chen, A. Nirula, B. Heller, R.L. Gottlieb, J. Boscia, J. Morris, G. Huhn, J. Cardona, B. Mocherla, V. Stosor, I. Shawa, A.C. Adams, J. Van Naarden, K.L. Custer, L. Shen, M. Durante, G. Oakley, A.E. Schade, J. Sabo, D.R. Patel, P. Klekotka, D.M. Skovronsky, BLAZE-1 investigators. SARS-CoV-2 neutralizing antibody LY-CoV555 in outpatients with covid-19, N. Engl. J. Med. 384 (3) (2021 Jan 21) 229–237, https://doi.org/10.1056/NEJMoa2029849. Epub 2020 Oct 28. PMID: 33113295; PMCID: PMC7646625.

[51] A. Maimaiti, Z. Feng, Y. Liu, M. Turhon, Z. Xie, Y. Baihetiyaer, X. Wang, M. Kasimu, L. Jiang, Y. Wang, Z. Wang, Y. Pei, N7-methylguanosin regulators-mediated methylation modification patterns and characterization of the immune microenvironment in lower-grade glioma, Eur. J. Med. Res. 28 (1) (2023 Mar 30) 144, https://doi.org/10.1186/s40001-023-01108-4. Erratum in: Eur J Med Res. 2024 Jan 19;29(1):59. PMID: 36998056; PMCID: PMC10061823.

[52] B. Yang, A. Gu, Y. Wu, High EIF4E2 expression is an independent prognostic risk factor for poor overall survival and recurrence-free survival in uveal melanoma, Exp. Eye Res. 206 (2021) 108558, https://doi.org/10.1016/j.exer.2021.108558.

[53] S. Evagelou, O. Bebenek, E. Specker, J. Uniacke, DEAD box protein family member DDX28 is a negative regulator of hypoxia-inducible factor 2α- and eukaryotic initiation factor 4e2-directed hypoxic translation, Mol. Cell Biol. (2020), https://doi.org/10.1128/MCB.00610-19.

[54] G. Melanson, S. Timpano, J. Uniacke, The eIF4E2-directed hypoxic cap-dependent translation machinery reveals novel therapeutic potential for cancer treatment, Oxid. Med. Cell. Longev. 2017 (2017) 6098107, https://doi.org/10.1155/2017/6098107.

[55] E.C. Little, E.R. Camp, C. Wang, P.M. Watson, D.K. Watson, D.J. Cole, The CaSm (LSm1) oncogene promotes transformation, chemoresistance and metastasis of pancreatic cancer cells, Oncogenesis 5 (2016) e182.

[56] P.M. Watson, S.W. Miller, M. Fraig, D.J. Cole, D.K. Watson, A.M. Boylan, CaSm (LSm-1) overexpression in lung cancer and mesothelioma is required for transformed phenotypes, Am. J. Respir. Cell Mol. Biol. 38 (2008) 671–678.

[57] L.U. Jing-shan, Jian Yang, X.U. Ying, Y.U. Chang-xi, Advances in application of metabolomics in rheumatoid arthritis research, Chin. Pharmacol. Bull. 35 (9) (2019) 1193–1196.

[58] W.A. Alasmari, A. Abdelfattah-Hassan, H.M. El-Ghazali, S.A. Abdo, D. Ibrahim, N.A. ElSawy, E.S. El-Shetry, A.A. Saleh, M.A.S. Abourehab, H. Mahfouz, Exosomes derived from BM-MSCs mitigate the development of chronic kidney damage post-menopause via interfering with fibrosis and apoptosis, Biomolecules 12 (5) (2022 May 2) 663, https://doi.org/10.3390/biom12050663. PMID: 35625591; PMCID: PMC9138582.