Systemic Lupus Erythematosus Detection with Machine Learning Models

Jishnu Guthikonda

Background

Systemic Lupus Erythematosus (SLE) is a chronic autoimmune disease characterized by loss of immune tolerance, production of autoantibodies, and inflammation affecting multiple organ systems [1]. Autoimmune diseases, including SLE, cause inflammation, pain, and damage in different organs and tissues [1]. SLE symptoms and outcomes vary drastically among patients challenging the understanding regarding the pathology. Hence, some people experience relatively mild symptoms, while others develop severe complications involving organs such as the kidneys, heart, lungs, or brain [1].

There are different types of lupus, including Systemic Lupus Erythematosus (SLE), Cutaneous Lupus Erythematosus (CLE), Neonatal Lupus, and Drug-Induced Lupus [4]. Of these, SLE is the most common and the most severe owing to the susceptibility of every organ system [1,4]. Interestingly, the incidence of lupus is more females than males, particularly during the reproductive years, with an approximate female-to-male ratio of 9:1 [1,2]. This difference suggests that sex hormones, including estrogen, contribute to disease susceptibility and immune dysregulation [1,2].

Emerging evidence suggests that SLE develops through a combination of genetic susceptibility, hormonal influences, environmental triggers, and immune dysfunction [1,2]. Environmental factors such as ultraviolet light exposure, infections, and smoking have been associated with disease onset or disease flares [1]. In patients with lupus, the immune system becomes overactive and produces autoantibodies that target the body's own tissues, resulting in chronic inflammation, immune complex formation, complement activation, and subsequent tissue damage [1,3].

The symptoms of SLE vary considerably between individuals. Common manifestations include fatigue, fever, weight loss, joint pain, skin rashes, and photosensitivity [1]. Additionally, lupus leads to nephritis, cardiovascular disease, pulmonary disease, hematologic abnormalities, or neurological complications [1]. A major challenge associated with lupus is a relapsing-remitting course, with periods of increased disease activity, known as flares, followed by periods of remission or reduced symptoms [1].

Diagnosis of lupus relies on evaluating both clinical manifestations and laboratory findings [1]. A commonly used screening test is the antinuclear antibody (ANA) test, which is highly sensitive for SLE; however, has been hurdled by poor specificity [1]. Biomarkers, including anti-double stranded DNA antibodies, anti-beta 2 glycoprotein I antibodies, complement proteins C3 and C4, lupus anticoagulant, and anti-Smith antibodies, provide greater diagnostic specificity reflecting the disease status [1,3].

Machine learning (ML) techniques have been increasingly applied to improve the diagnosis of SLE [6–9]. Generally, ML models use laboratory results, gene expression data, or electronic health record information to identify patients with SLE or predict disease progression and flares

[6–9]. Algorithms including logistic regression, random forests, support vector machines, gradient boosting methods, and neural networks have demonstrated promising diagnostic performance in multiple studies [6–9]. However, systematic reviews have noted that the published models remain limited by relatively small sample sizes, insufficient external validation, risk of bias, and limited clinical applicability, highlighting the need for models that are more robust and generalizable [6,7].

In this study, we attempted to create an accurate machine learning model capable of supporting the clinical diagnosis of SLE. This is significant because SLE remains one of the most difficult autoimmune diseases to diagnose due to its highly variable clinical presentation, heterogeneous biomarker profiles, and overlap with other autoimmune diseases [1,6]

Methods:

A machine learning model was built on publicly available data containing the randomly generated lab results of synthetic patients who were randomly generated to have an autoimmune disease. The chosen dataset contained over 10,000 entries of synthetic patients with different autoimmune diseases. The dataset contained a ratio of 9907*:93* (SLE to Non-SLE patients). The 2019 EULAR/ACR criteria for diagnosing SLE was employed to select relevant parameters from the dataset on which to build the model [1,3]. The parameters include age, gender, sickness duration months, hemoglobin, WBC count, PLT count, ANA, C3, C4, Anti-dsDNA, Anti-Sm, low-grade fever, stiffness in the joints, and joint pain [1,3]. Binary Cross Entropy (BCE) Loss function was employed to help the model learn how to take in data and output an accurate result. BCE loss is a function used to measure the difference between the model's predicted probability and the actual outcome in binary classification problems. Since the model was designed to determine whether a patient had SLE or not, BCE loss was an appropriate choice because it penalizes incorrect predictions and rewards predictions that are closer to the correct diagnosis. During training, the model continuously adjusted its parameters to minimize the BCE loss, allowing it to improve its ability to accurately classify patients based on their clinical and laboratory data. Gemini was employed to aid with data processing of the original dataset. Copilot was used to help auto generate code to increase the efficiency of the coding. ChatGPT was used to generate the plots.

The dataset was randomly divided into 70% training data and 30% testing data. The training set was used to train the model, while the testing set was used to evaluate its performance on unseen data. Before training, the features were standardized using the mean and standard deviation of the training set, and the same scaling parameters were applied to the testing set to prevent data leakage.

K-Fold model description

Repeated random train–test splits were performed using different random seeds to validate the K-Fold model. Model performance was evaluated on the testing set after each run, and the accuracy from each trial was recorded. The mean accuracy across all runs was then calculated

to provide a more reliable estimate of the model's overall performance and reduce the effect of variation from any single data split [6,7].

K-Fold Model Validation

A 10-fold cross-validation approach was used to evaluate the model's performance and generalizability. The dataset was randomly shuffled and divided into ten equal folds, with nine folds used for training and one fold used for testing during each iteration until every fold had served as the test set once. Feature scaling was performed using only the training data in each fold to prevent data leakage, and the model's accuracy, recall, precision, and F1-score were averaged across all ten folds. This method provided a more reliable estimate of performance by reducing the impact of any single train-test split.

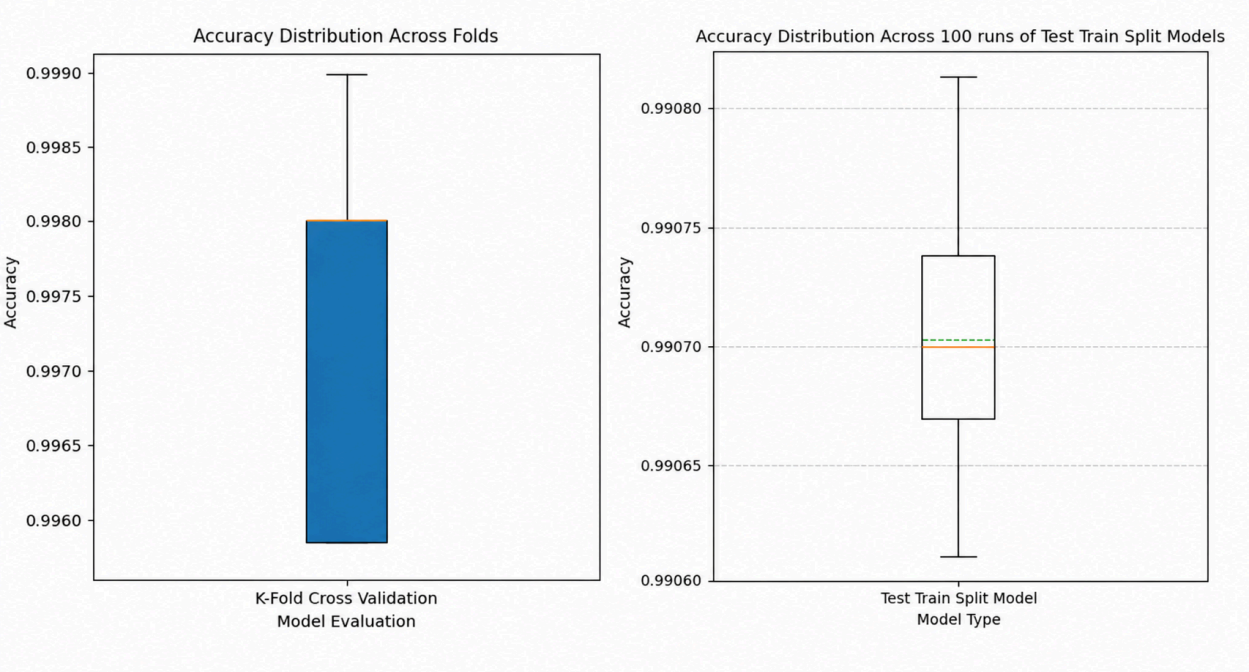
Statistical analysis:
Model performance was evaluated using descriptive classification metrics rather than inferential statistical tests. Accuracy, precision, recall (sensitivity), specificity, and F1 score were calculated for both the train-test split and 10-fold cross-validation models, and a Precision-Recall curve was generated to assess performance across different decision thresholds. These analyses were used to evaluate the model's predictive accuracy, reliability, and ability to distinguish SLE cases from non-SLE cases.

Results:

The test train split model was evaluated using standard classification metrics, including accuracy, precision, recall, and F1 score. It achieved an accuracy of 0.9973, indicating that the model correctly classified the vast majority of cases overall. However, additional metrics were used to better understand performance beyond overall correctness. The sensitivity of the model was 0.8515, showing that when the model predicted a positive diagnosis for SLE, it was correct approximately 85.2% of the time. The specificity was 0.9990, meaning the model labeled a healthy patient as healthy 99.9% of the time The precision was 0.8859, indicating that 88.6% of the model’s predicted SLE cases were true cases. The F1 score, which balances precision and recall, was 0.8535, reflecting a strong but slightly imbalanced trade-off between correctly identifying SLE cases and minimizing false positives in the test train split model.

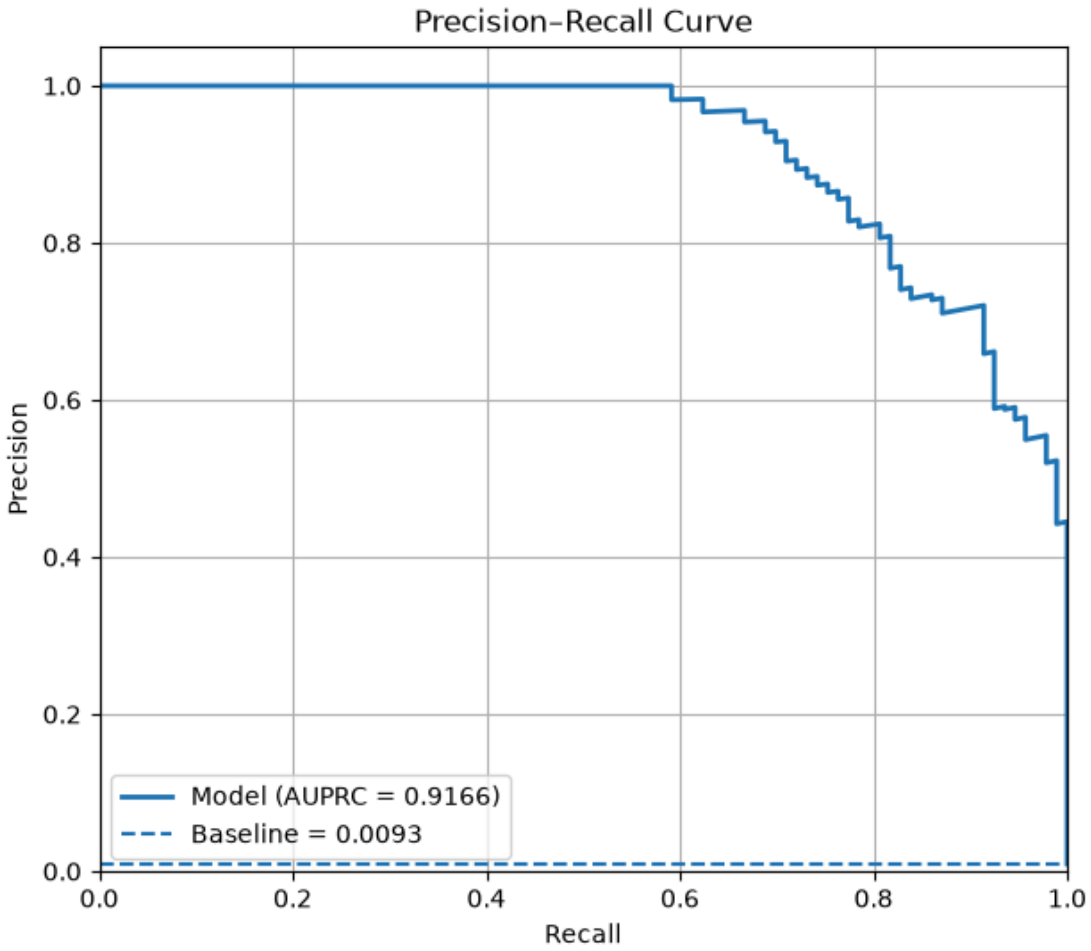
The k-fold cross-validation model was evaluated using standard classification metrics, including accuracy, precision, recall, and F1 score. It achieved an average accuracy of 0.9974 across the 10 folds, indicating that the model correctly classified the vast majority of patient cases overall. However, additional metrics were used to better understand performance beyond overall accuracy, especially due to the large imbalance between SLE and non-SLE cases in the dataset. The sensitivity of the model was 0.9153, showing that it correctly identified approximately 91.5% of actual SLE cases. The specificity was 0.9983, meaning that the model correctly classified non-SLE patients as healthy approximately 99.83% of the time. The precision was 0.8340, indicating that 83.4% of the model’s predicted SLE cases were true SLE diagnoses. The F1 score, which balances precision and recall, was 0.8625, demonstrating a strong balance between identifying SLE cases and limiting false positive predictions across the 10-fold cross-validation model.

Overall, the high accuracy suggests strong general predictive performance on the dataset, while the precision, recall, and F1 score provide a more nuanced view of the model's ability to correctly detect SLE cases without overpredicting positive diagnoses (**Figure 1**).



The distribution of model accuracy was evaluated using both 10-fold cross-validation and 100 independent train-test split runs, as shown in Figure 1. The K-fold cross-validation model using 10 folds, achieved consistently high accuracy, with values ranging from approximately 99.6% to 99.9%, demonstrating strong performance across different subsets of the dataset. Similarly, the 100 train-test split models produced highly consistent results, with a median accuracy of approximately 99.07% and only a narrow spread between runs. The small interquartile ranges and short whiskers observed in both boxplots indicate low variability in model performance, suggesting that the classifier generalizes well and is not overly dependent on a particular partition of the data. Overall, the similarity between the two evaluation methods demonstrates that the model is both accurate and stable across multiple validation strategies.

Figure 2

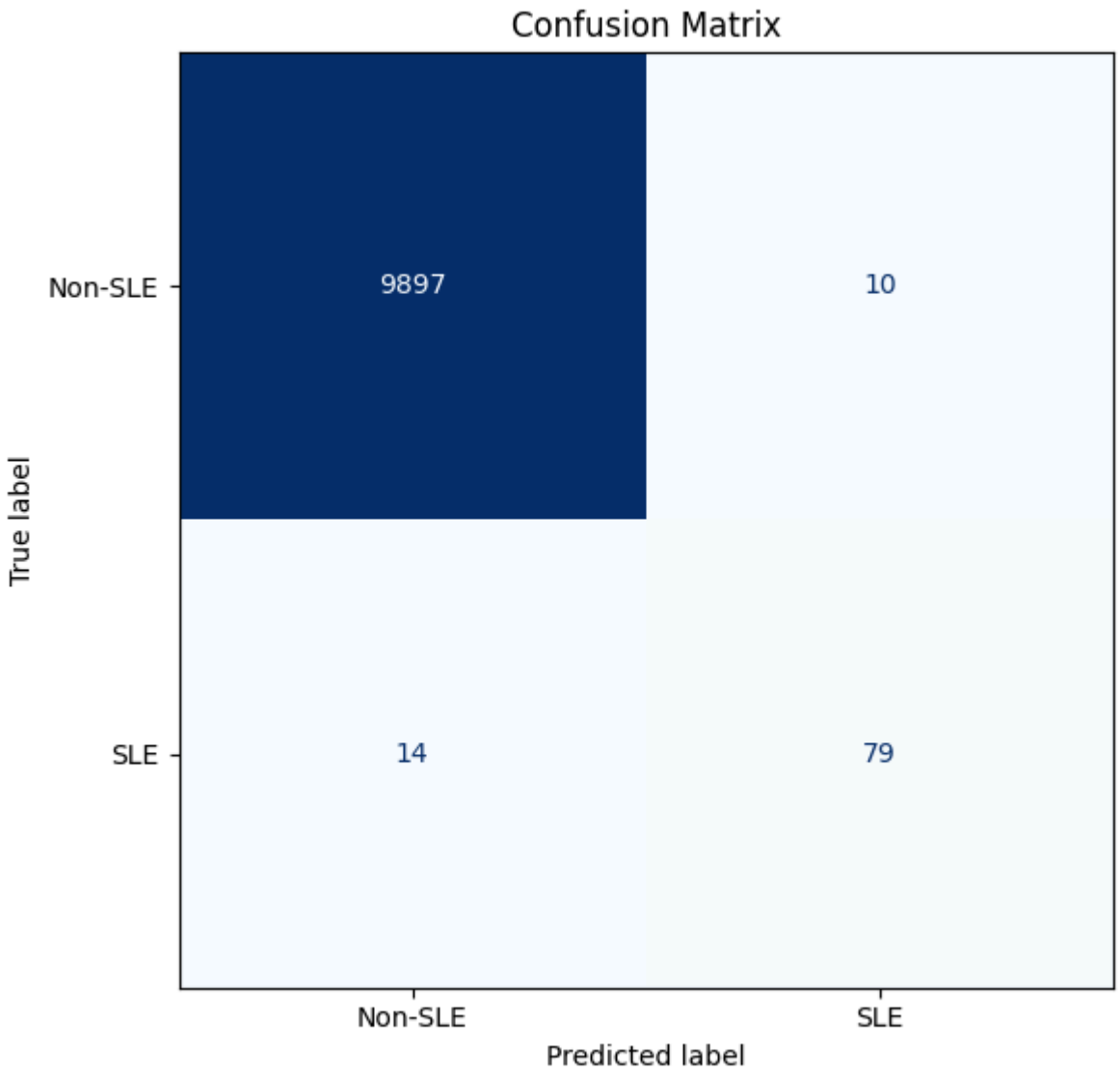


To evaluate the performance of the trained model, a Precision–Recall analysis was conducted by plotting precision as a function of recall across a range of decision thresholds. As shown in Figure 2, the model achieved an Area Under the Precision–Recall Curve (AUPRC) of 0.9166, indicating excellent performance in distinguishing positive cases despite the highly imbalanced dataset. As shown in the curve, the model maintains perfect precision over the initial portion of the graph while recall increases to nearly 0.60, demonstrating that the highest-confidence predictions are made with virtually no false positives. Beyond this point, precision decreases gradually as recall continues to improve, reflecting the expected trade-off between identifying additional true positive cases and introducing a small number of false positives. Even at high recall values approaching 1.0, the model retains moderate precision, substantially outperforming the baseline prevalence of 0.0093. Overall, these results indicate that the model effectively balances sensitivity and precision, making it well suited for detecting Systemic Lupus Erythematosus while minimizing unnecessary positive classifications.

Discussion:

The results of this study demonstrate that the machine learning model was able to accurately identify patients with systemic lupus erythematosus (SLE) using commonly collected clinical and laboratory data. The model achieved an overall accuracy of 99.73%, along with a precision of 88.59%, recall of 85.15%, F1 score of 85.35%, and an Area Under the Precision-Recall Curve (AUPRC) of 0.9166.

Figure 3



While the high accuracy shows that the model correctly classified nearly all patients, accuracy

alone is not enough to evaluate medical classification models because the dataset may contain more patients without SLE than with the disease. For this reason, precision, recall, and the F1 score provide a better understanding of the model's diagnostic ability. The high precision indicates that most patients predicted to have SLE actually had the disease, reducing unnecessary follow-up testing and treatment. Likewise, the recall shows that the model successfully identified most true SLE cases, although a small number of patients were still missed. The F1 score demonstrates that the model maintains a strong balance between correctly identifying patients with SLE while limiting false positive predictions.

The model also showed strong reliability throughout testing. Both the 10-fold cross-validation and the 100 independent train-test split evaluations produced nearly identical accuracy distributions with very little variation between runs. This suggests that the model performs consistently regardless of how the data are divided and is not overly dependent on a single train-test split. In addition, the high AUPRC indicates that the model maintains strong predictive performance across different decision thresholds. The shape of the Precision-Recall curve shows that the model is able to identify many true positive cases while keeping false positive predictions relatively low. This is especially important for diseases such as SLE, where early diagnosis can lead to earlier treatment and better long-term patient outcomes.

One of the strengths of this study is that the model was built using variables selected from the 2019 EULAR/ACR classification criteria, making the model more applicable to real clinical practice. Since the input variables are commonly collected during routine laboratory testing, the model could potentially be incorporated into existing diagnostic workflows without requiring additional specialized testing. In addition, the large publicly available dataset containing over 12,500 patient records allowed the model to learn from a wide variety of autoimmune disease cases, helping improve its overall performance and stability.

Despite these promising results, several limitations should be considered. The model was trained and tested using only one publicly available dataset, so additional testing with independent datasets from different hospitals and patient populations is needed before the model could be used in clinical settings. Furthermore, the dataset did not include potentially useful information such as family history, medical history, imaging results, or changes in laboratory values over time. Including these variables in future models may improve the model's ability to detect additional SLE cases while maintaining high precision. Future studies could also investigate other machine learning algorithms and explainable artificial intelligence techniques to better understand which clinical features contribute most to each prediction.

Overall, the findings of this study suggest that machine learning can be an effective tool for supporting the diagnosis of systemic lupus erythematosus. The model demonstrated high predictive performance, consistent results across multiple validation methods, and strong discrimination between patients with and without SLE. Although further validation is necessary before clinical implementation, this study shows that machine learning has the potential to assist physicians in making earlier and more accurate diagnoses, ultimately improving patient care and outcomes.

Conclusion:
The results of this study demonstrate that machine learning can be successfully applied to the identification of Systemic Lupus Erythematosus (SLE) using clinical and laboratory data from patients suspected of having an autoimmune disease. After training on a dataset containing over 12,500 patient records, the model achieved an accuracy of 99.73%, indicating that it was able to correctly classify the vast majority of patient cases. Additional evaluation metrics were used to provide a more complete assessment of model performance. The model achieved a precision of 88.59%, meaning that when the model predicted a patient had SLE, it was correct approximately 88.6% of the time. The model also achieved a recall of 85.15%, demonstrating its ability to successfully identify most patients who truly had SLE. Furthermore, the model obtained an F1 score of 85.35%, indicating a strong balance between precision and recall.

Because the primary purpose of the model was to identify patients with SLE, recall was considered an especially important metric. The model's ability to correctly identify over 85% of true SLE cases suggests that machine learning may serve as a useful tool for assisting healthcare providers in the diagnostic process. While the model does not replace clinical judgment, it has the potential to support earlier detection of SLE by identifying patients who may benefit from additional testing and evaluation. Overall, the results suggest that machine learning models trained on clinically relevant laboratory and symptom data can effectively distinguish patients with SLE from those with other autoimmune diseases and may contribute to improved diagnostic decision-making in the future.

Bibliography

1. Bertsias, G., & Mavragani, C. P. (2025). *Systemic lupus erythematosus*. In *StatPearls*. StatPearls Publishing. https://www.ncbi.nlm.nih.gov/books/NBK535405/
2. Online Mendelian Inheritance in Man (OMIM). (n.d.). *Systemic lupus erythematosus (MIM No. 152700)*. Johns Hopkins University. https://www.omim.org/entry/152700
3. Song, Y., & Kono, M. (2025). Potential biomarkers in systemic lupus erythematosus. *JMA Journal, 8*(3), 689–698. https://doi.org/10.31662/jmaj.2025-0190
4. Cojocaru, M., Cojocaru, I. M., Silosi, I., & Vrabie, C. D. (2011). Manifestations of systemic lupus erythematosus. *Maedica, 6*(4), 330–336. https://pmc.ncbi.nlm.nih.gov/articles/PMC3351863/
5. Fanouriakis, A., Tziolos, N., Bertsias, G., & Boumpas, D. T. (2022). Systemic lupus erythematosus: Classification, diagnosis, and clinical management. *Journal of Clinical Medicine, 11*(17), 5063. https://doi.org/10.3390/jcm11175063
6. Wang, Z., et al. (2025). Validity and applicability of machine learning models for systemic lupus erythematosus diagnosis. *Lupus Science & Medicine, 13*(1), e002007. https://lupus.bmj.com/content/13/1/e002007
7. Zhou, Y., Wang, M., Zhao, S., & Yan, Y. (2022). *Machine learning for diagnosis of systemic lupus erythematosus: A systematic review and meta-analysis*. *Computational Intelligence and Neuroscience, 2022*, Article 7167066. https://doi.org/10.1155/2022/7167066 *(Retracted in 2023.)*
8. Park, M. (2024). Improving the diagnosis of systemic lupus erythematosus with machine learning algorithms based on real-world data. *Mathematics, 12*(18), 2849. https://doi.org/10.3390/math12182849

9. Bortoluzzi, A., et al. (2024). A machine learning model for systemic lupus erythematosus diagnosis based on routine laboratory tests. *Clinical and Experimental Rheumatology*. https://www.clinexprheumatol.org/article.asp?a=19813