SUBMITTED 23 JUN 22
REVISIONS REQ. 18 SEPT & 14 NOV 22; REVISIONS RECD. 23 OCT & 24 NOV 22
ACCEPTED 8 DEC 22
ONLINE-FIRST: DECEMBER 2022
DOI: https://doi.org/10.18295/squmj.12.2022.069

Machine Learning Approach for Predicting Systemic Lupus Erythematosus in Oman-based Cohort

***AlHassan AlShareedah,[1] Hamza Zidoum,[1] Sumaya Al-Sawafi,[1] Batool Al-Lawati,[2] Aliya Al-Ansari[3]**

Departments of [1]Computer Science and [3]Biology, College of Science and [2]Department of Medicine, College of Medicine, Sultan Qaboos University, Muscat, Oman.
**Corresponding Author's e-mail: al.hassan.satii@gmail.com*

Abstract

Objectives: Design a machine learning-based prediction framework to predict the presence or absence of Systemic Lupus Erythematosus (SLE) in a cohort of Omani patients. ***Methods:*** Records of 219 patients from 2006 to 2019 were extracted from SQU Hospital electronic records, 138 patients have SLE, and the remaining 81 have other rheumatologic diseases. Clinical and demographic features were analyzed to focus on the early stages of the disease. Our design implements Recursive Feature Selection (RFE) to select only the most informative features. In addition, the CatBoost classification algorithm is utilized to predict SLE and an explainer algorithm (SHAP) is applied on top of the CatBoost model to provide individual prediction reasoning which is then validated by rheumatologists. ***Results:*** CatBoost achieved an Area Under the ROC curve (AUC) score of 0.95 and a Sensitivity of 92%. Four clinical features (Alopecia, renal disorders, Acute Cutaneous Lupus, and hemolytic anemia) along with the patient's age were shown to have the greatest contribution to the prediction by the SHAP algorithm. ***Conclusion:*** We have designed and validated an explainable framework to predict SLE patients and provide reasoning for its prediction. Our framework enables early intervention for clinicians which leads to positive healthcare outcomes.

Keywords: Systemic Lupus Erythematosus; Interpretation; Machine Learning; Supervised; Clinical Decision Support System; Statistical Data; Data Analysis.

Advances in Knowledge

- The first self-explainable prediction framework for SLE disease specific to the Omani population is developed.
- Achieved an AUC score of 0.956 and Sensitivity of 92%.
- Identifies patterns in clinical manifestation which are unique to the Omani population.
- The patient's age and four clinical features (renal disorders, alopecia, cutaneous lupus, and hemolytic anemia) had the highest contribution to the model's prediction.
- Compared to other Arab ethnicities, renal disorders frequency in Oman was the highest while alopecia frequency was the lowest.
-

Application to Patient Care

- The model can potentially be used as a clinical decision support system that alerts clinicians to the presence of SLE which prompts further investigation until an official diagnosis is made.
- Enabling clinicians to contrast the information reported by the model with their knowledge through an interpretation algorithm. Thereby increasing the probability of correct diagnosis and encouraging the adoption of Machine Learning (ML) in healthcare.
- A practical introduction of machine learning and interpretation tools to the medical diagnosing process that improves early detection of SLE; a crucial factor in lowering flare rate and reducing mortality.

Introduction

Systemic lupus erythematosus (SLE) is a chronic multisystem autoimmune disease. SLE is caused by genetic and environmental factors that potentiate the creation of high-titer autoantibodies aimed at native DNA and other cellular elements.[1] The creation of these autoantibodies leads to a pathological process that manifests into different medical conditions in different organ systems, from skin arthralgia to cardiovascular and renal morbidity.[2] The clinical phenotype of SLE varies with race, gender, and age which makes the disease difficult

to diagnose.[3] In Oman, it is estimated that the mortality rate is at 5% and the mean prevalence is 38 per 100,000 individuals,[4] this is higher than in Saudi Arabia and lower than in UAE. Initial SLE symptoms are often nonspecific and mimic other medical conditions, increasing the risks of diagnostic delay. Additionally, the heterogeneity of manifestations makes early diagnosis even more difficult and subsequently delays the start of effective treatment before the occurrence of organ damage.

In recent years, great improvements in treatment strategies for SLE have been made. However, despite the improved prognosis, various challenges remain for the diagnosis and therapeutic management of SLE.[5] One of those challenges is early diagnosis. SLE onset is gradual and clinically-evident manifestation develops over the years. Moreover, a variety of conditions may mimic SLE conditions, including infectious and hematologic diseases.[6] It has been proven from database analysis that patients with a diagnosis window below 6 months (between probable SLE onset and diagnosis) had low flare rates and hospitalizations compared with patients with late diagnosis.[7] Late diagnosis is also associated with the risk of developing progressive organ damage and subsequently increases the mortality rate.[8]

This study focuses on effective SLE prediction as well as finding the associated clinical features. With the aid of interpretation tools, clinicians can understand the decision-making process of Machine Learning (ML) models. This, in turn, enables clinicians to be alerted to different manifestations and symptoms at early stages and provide better healthcare outcomes. The model is trained on a local cohort of 219 Omani patients with SLE as well as other control diseases. Additionally, we identified the minimum set of clinical and demographic features required for an accurate prediction. Finally, an explainable approach based on SHapley Additive exPlanations (SHAP) method was applied to generate individual explanations of the model's decisions as well as ranking clinical features by contribution.

Methods

The dataset used in this study was collected from structured and unstructured sources. This includes the Electronic Medical Records (EMR) in Sultan Qaboos University Hospital's Rheumatology clinic named TrakCare. TrakCare stores the patients' information, medical

state, and medical history. Patients' demographic data were obtained directly from TrakCare meanwhile clinical data was unstructured as it was stored in the patient's medical history as clinical notes from each visit to the hospital. Entry criteria for Rheumatology patients is a positive Antinuclear Antibodies test (ANA test) while the Exclusion criteria included all non-Omani patients as well as patients with non-sufficient data. To separate patients with SLE and control diseases, the most recent SLE classification criteria set by EULAR/ACR were used.[9] When applied, patients with a score of 10 or above are diagnosed with SLE. A total of 219 patient records match the entry criteria, 138 are diagnosed with SLE, and 81 have other control diseases, this was also validated by a rheumatologist on case-by-case bases.

Our framework contains three main stages, starting with feature selection that reduces noisy data and utilizes only the most informative features followed by the classifier, which trains and tests the model to predict the presence of SLE. After the model is trained, the explainer algorithm proceeds to provide individual prediction breakdown through informative visual plots.

In the first stage [Figure 1], the recursive feature elimination (RFE) algorithm with ten-fold cross-validation (CV) was used. RFE works by building a model, selecting the best feature, picking out the selected feature, and then repeating this process for the remaining features until all the features are traversed. For the second stage of this framework, we have implemented Categorical Boosting or CatBoost, an ensemble learning algorithm that is based on gradient boosting.[10]

For the final stage, the SHAP library is implemented.[11] SHAP calculates ‘Shapley values’ for each feature to determine the contribution of a feature to the final prediction represented by the magnitude and sign of the Shapley value. Specifically, the importance of the feature relative to the prediction is represented by the magnitude of the Shapley value. SHAP tool can also perform local and global interpretability simultaneously. With the help of SHAP algorithm, we can break down each prediction individually. As a demonstration, we took two individuals from the testing set, one that was predicted to have the disease and one that was not. Three types of figures were used to show the prediction breakdown, force plot, waterfall plot, and

summary plot. The force plot demonstrates how the features contributed to the model's prediction for a specific observation. The colors in the force plot correspond to the feature pushing the prediction probability higher or lower. The target in our model has two classes, class 1 for a positive diagnosis of SLE and class 0 for a negative diagnosis of SLE. To obtain a full list of features ranked by their contribution we use a waterfall plot. The summary plot displays the feature's effects and their importance. Each point on the summary plot represents a Shapley value for a feature and an instance.

To train and validate the performance of CatBoost, the dataset was divided into training and testing sets. The former is used to train the model and the latter is used to test the performance of the model. Additionally, a subset of the training data set was used for cross-validation to protect the models from overfitting and optimize the model's parameters. Each of the models undergoes a hyper-parameter optimization through grid search with five-fold cross-validation. To avoid reporting biased results and limit overfitting, we calculated the measurement's average of 10 repetitions for each model. Finally, three other classifiers were evaluated similarly, which are Multi-Layer Perceptron (MLP), Support Vector Machine (SVM), and Random Forest. Their performance evaluations were compared to CatBoost to observe the effectiveness of CatBoost. The classifiers were selected based on related studies that employed ML for disease prediction.[12 13]

Due to the imbalanced nature of the problem, the AUC (area under ROC curve) and Sensitivity parameters are used to evaluate the classification performance.

The study was approved by the Ethics Committee of the College of Medicine and Health Science at Sultan Qaboos University (SQU) in protocol number MERC #1418 and #1650. No participant consent is required for this study as per the regulation of Sultan Qaboos University's Hospital.

Results

The extracted data covers patient records from January 2006 to December 2019. Female patients represent the majority of our records with 92%. Patients between 25 years old and late

30's represent the largest age group with a mean age of 38. Al Batinah Governorate had the highest number of patients (37.9%) followed by Muscat (23.7%).

Initial data contained 28 clinical, demographic, laboratory variables (so-called "features" in ML), and no missing values were found in the data [Table 1]. Laboratory features include immunological test results such as Anti-dsDNA Test, Anti-Smith antibody, and more. These features however, are highly sensitive to SLE and can introduce bias to the prediction model therefore it was dropped. The remaining data consist of 20 clinical and demographic features. The majority of the features are represented by non-numerical (categorical) values. This entails a transformation (encoding) to numerical values as this is a prerequisite for all statistical models. Thus, Ordinal encoding was applied, moreover, because of the variance in range for different features, Min-Max normalization was also applied.[14]

Applying the RFE feature selection algorithm, the optimal number of features selected was 13. From the RFE selected features, three demographic features, as well as 10 clinical features, were selected. CatBoost had an AUC score of 0.956, with the Random Forest classifier and SVM scoring 0.935 and 0.916 AUC respectively. For Sensitivity, CatBoost had 92%, Random Forest achieved 89% and SVM score is 86%.

Two samples from the testing set were used to generate the different SHAP plots. The first sample (Patient 1) is predicted to have SLE, the force plot attributes this to renal disorders, and the patient's age [Figure 2.a]. Since the values are normalized we cross-referenced them with test data and found that the patient's age is 40 which falls within the age group SLE that is most active. Additionally, the patient has been diagnosed with Lupus Nephritis a disease that is commonly caused by an auto-immune disorder. On the other hand, the second sample (patient 2) displays a lack of any autoimmune manifestation [Figure 2.b], a long disease duration, and the age of 56 makes him outside the age group that SLE is most active.

Looking at the waterfall plot for patient 1[Figure 3.a], the feature with the highest SHAP value is Renal by a large margin. Due to its high SHAP value, the presence of renal disorder in patient 1 had the greatest contribution to the positive prediction of SLE. This was followed by

the age and province features. Overall, four blue features were pushing the prediction probability lower toward class 0. The non-existence of alopecia, hemolytic anemia, and Acute Cutaneous Lupus (ACL) in the patient 1 profile resulted in negative SHAP values. The remaining features had minimal impact on the prediction probability evidenced by their low SHAP values. In contrast, the waterfall plot for patient 2[Figure 3.b] indicates that age is the largest contribution toward class 0, followed by the absence of any renal disorders.

In [Figure 4], similar to what was deduced, the older the patient is the less likely it is to have SLE, which is evident by the red dots on the negative scale of SHAP values. The same can be said for disease duration, we find that long disease durations without autoimmune manifestation correlated with the absence of SLE. Our result indicates that the higher the patient's age and disease duration the less likely that SLE is the cause. Renal disorders are ranked the highest in contribution followed by alopecia, ACL, and hemolytic anemia. The lowest contributing features are Serositis, Proteinuria, and Leukopenia.

Discussion

In clinical applications, the ability to justify the prediction is equally as important as the prediction score itself. This is because of the high sensitivity of the medical environment where misclassification could lead to devastating results. It is therefore challenging to trust complex ML models for several reasons. First, the models are often designed and rigorously trained on specific diseases in a narrow environment. Second, it depends on the user's technical knowledge of statistics and ML. Third, how the data is labeled affects the results produced by the model.[15] For these reasons and more, Interpretable ML has thus emerged as an area of research that aims to design transparent and explainable models by developing means to transform a black-box ML model into white-box ML models. By providing transparent prediction, domain experts can accurately interpret the results meaningfully.

Through the use of SHAP algorithm, clinicians can understand the model's reasoning, thus resembling clinical reasoning. Our model is situated between early to mid-screening suggesting that physicians have minimum visible clinical symptoms and subsequently no immunological test data.[16] The model can reasonably make predictions that can alert clinicians

to investigate the presence of SLE by requesting immunological tests once suspicion of SLE is predicted. Specifically, the ANA test and the anti-double stranded DNA (anti-dsDNA) are highly sensitive and decisive if found positive.[17] Additionally, an immunologist has compared multiple individual prediction breakdown plots and validated the results and the model justification.

One of the features that were used to profile patients are age, age-onset, and disease duration features. It was deduced from the SHAP algorithm that older patients were the least affected by the disease. Similarly, patients with long disease duration without adverse manifestations such as anemia or lupus nephritis are shown statistically to be less likely diagnosed with SLE. Experts point out, however, that SLE intensity increases and decreases at intervals differently from patient to patient, thus in rare occasions clinical symptoms might not manifest until the late phases of the disease.[18] Research suggests that late-onset SLE occurs at a rate of 3-18% in the exposed population.[19]

Renal disorders were the highest feature in contribution according to SHAP [Figure 4]. This was in concordance with Beckwith and Lightstone (2014) who states that about 40-70% of SLE patients develop clinically diagnosed renal involvement which is known as Lupus Nephritis.[20] Lupus Nephritis (LN) is commonly diagnosed through kidney biopsy, previous research identified proteinuria, urine protein-to-creatinine ratio, anti-dsDNA, and complement levels as laboratory markers of LN. However, these LN laboratory markers lack specificity and sensitivity for identifying renal activity and damage.[21] In Oman, LN is the most frequent glomerular disease occurring in about 30%-36% of all patients who had a renal biopsy. This is supported by Al Adhoubi (2020),[4] where 52% of SLE patients have developed LN. Despite the majority of our data lacking kidney biopsy information, LN is also present in 11% of patients with renal disorders.

Moreover, we found other clinical features that had about the same influence on the prediction. These are Alopecia, Cutaneous Lupus, and Anemia. Alopecia is a hair loss that also varies in damage activity from non-scarring to scarring. Currently, it is estimated that more than half of SLE patients develop alopecia, although most of the research that estimates alopecia

prevalence is limited by the small population size. Cutaneous Lupus which includes a butterfly rash across the face between the eyes and nose. ACL is a sign of VGLL-3 & Anti-SSA antibodies which indicate skin damage activity caused by Lupus.[22] Anemia is the most common blood disorder, affecting about half of all people with active lupus.[23] Anemia is caused by a shortage of healthy red blood cells needed by the body to carry oxygen to the body's tissues. Hemolytic anemia, however, is not exclusive to SLE.

The prevalence of these influential clinical features across other Arab ethnicities was also investigated. While no study examined the differences between ethnicities within the Arab region, there have been few studies that have collected data on the SLE population locally. We looked at three cohorts from Saudi Arabia,[24] UAE,[25] and Egypt [Figure 5].[26] ACL or skin rash was found more prevalent in all other Arab cohorts reaching as high as 62% in UAE. Hemolytic anemia was the most varying feature, in Egypt and UAE, it is less prevalent than in Oman while in Saudi Arabia it is more prevalent than in Oman.[27] Renal disorders remained high at around 50% of all cohorts having some renal damage except for a slight decrease to 33% in Egypt. Studies also indicate that out of all renal biopsies, approximately 10%–36% are diagnosed with LN in the Gulf region. LN also tends to run a severe course in gulf populations with a high incidence of Class IV LN.[28]

Overall, with three critical features out of four found more prevalent in other Arab ethnicities, our model can be extended to include not only Omanis but also other Arab cohorts. It is important to note that all of these clinical features are not exclusive to SLE, but to autoimmune diseases in general. However, classification models can be trained to detect patterns specific to the Omani population, these patterns are the bases of the model's prediction for SLE presence.

These findings help to identify patterns in clinical manifestation which are unique to the Omani population and the Arab region by employing explainable prediction. Moreover, our research also highlights CatBoost algorithm, which had widespread attention in recent years for its fast calculation speed, powerful generalization ability, and strong predictive performance.[29] [30] [31] We achieved a margin of improvement of 0.21 AUC over the other classifiers, this may be attributed to its novel implementation of ordered boosting, and

permutation-driven alternative to the classic algorithm. This study also acknowledges the problem with imbalanced classification evaluation where the research is biased toward the performance of cases that are poorly represented in the data samples.[32] Standard evaluation criteria tend to focus the evaluation of the models on the most frequent cases, thus if applied, could lead to sub-optimal classification models. Thus, AUC and Sensitivity were selected as evaluation criteria.

Finally, by combining the framework's prediction with the interpretation algorithm we are promoting self-explainable frameworks that enable physicians to make meaningful decisions based on ML-based information combined with their knowledge. Thereby improving the probability of correct diagnosis and encouraging the adoption of ML in healthcare. These goals however are hindered by the retrospective nature of the data. An ideal framework is much more effective with longitudinal data of SLE patients that include pre-diagnosis profiles before the appearance of adverse symptoms. Moreover, our framework may not scale properly with large datasets. Specifically, large data will significantly increase the computational time for SHAP, and categorical data with high cardinality is inefficient with the Ordinal encoder algorithm.[33] Different tools can also be applied to increase the accessibility and presentation of our model such as presenting the outcome as a prediction probability instead of a binary value.

Conclusion

This study proposes a three-stage interpretable framework for predicting the presence or absence of SLE in an Omani cohort of 219 patients. CatBoost classifier and SHAP interpretation tool were implemented to predict and justify individual predictions and eliminate any risk of misclassification. Four clinical features were identified to have the highest influence on the prediction in addition to the patient's age. Alopecia, Renal, Acute cutaneous lupus, and Hemolytic Anemia are all indicators of lupus activity at varying rates, combined with the patient's age and age-onset the model was able to establish a profile of the disease relative to Omanis. Overall, our findings aid in providing a practical introduction of machine learning and interpretation tools to medical diagnosis, thereby increasing the efficiency of medical testing and subsequently enabling early intervention which leads to better treatment and a positive healthcare outcome.

Authors' Contribution

HZ and AA conceived the idea. H.Z. and A.A. designed the study. SA collected the data. BA and AA-A have validated the data and results. Research experiments, implementation, and results were performed by AA with input from HA. AA drafted the manuscript with edits from HZ, AA-A and BA. All authors approved the final version of the manuscript.

Acknowledgment

We would like to thank the Oman Research Council for supporting this research. We also gratefully acknowledge the support from the Sultan Qaboos University Hospital for their cooperation.

Conflict of Interest

The authors declare no conflicts of interest.

Funding

This work was supported in part by an internal Grant.

References

1. Gaipl U, Munoz L, Grossmayer G, Lauber K, Franz S, Sarter K et al. Clearance deficiency and systemic lupus erythematosus (SLE). J Autoimmun 2007;28(2-3):114-121. doi: 10.1016/j.jaut.2007.02.005
2. Nisengard R. Diagnosis of systemic lupus erythematosus. Importance of antinuclear antibody titers and peripheral staining patterns. Arch Dermatol 1975;111(10):1298-1300. doi: 10.1001/archderm.111.10.1298
3. Lewis M, Jawad A. The effect of ethnicity and genetic ancestry on the epidemiology, clinical features and outcome of systemic lupus erythematosus. Rheumatology (Oxford) 2016; kew399. doi: 10.1093/rheumatology/kew399
4. Al-Adhoubi N, Al-Balushi F, Al Salmi I, Ali M, Al Lawati T, Al Lawati B et al. A multicenter longitudinal study of the prevalence and mortality rate of systemic lupus erythematosus patients in Oman: Oman Lupus Study. Int J Rheum Dis 2021;24(6):847-854. doi: 10.1111/1756-185 X.14130

5. Felten R, Lipsker D, Sibilia J, Chasset F, Arnaud L. The history of lupus throughout the ages. J Am Acad Dermatol 2020; doi: 10.1016/j.jaad.2020.04.150
6. Piga M, Arnaud L. The Main Challenges in Systemic Lupus Erythematosus: Where Do We Stand?. J Clin Med 2021;10(2):243. doi: 10.3390/jcm10020243
7. Oglesby A, Korves C, Laliberté F, Dennis G, Rao S, Suthoff E et al. Impact of Early Versus Late Systemic Lupus Erythematosus Diagnosis on Clinical and Economic Outcomes. Appl Health Econ Health Policy 2014;12(2):179-190. doi: 10.1007/s40258-014-0085-x
8. Murimi-Worstell I, Lin D, Nab H, Kan H, Onasanya O, Tierce J et al. Association between organ damage and mortality in systemic lupus erythematosus: a systematic review and meta-analysis. BMJ Open 2020;10(5):e031850. doi: 10.1136/bmjopen-2019-031850
9. Aringer M, Brinks R, Dörner T, Daikh D, Mosca M, Ramsey-Goldman R et al. European League Against Rheumatism (EULAR)/American College of Rheumatology (ACR) SLE classification criteria item performance. Ann Rheum Dis 2021;80(6):775-781. doi: 10.1136/annrheumdis-2020-219373
10. Dorogush A, Ershov V, Gulin A. CatBoost: gradient boosting with categorical features support. arXiv [Pre-print] 2018. doi: https://doi.org/10.48550/arXiv.1810.11363
11. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst 2017;30. doi: https://dl.acm.org/doi/10.5555/3295222.3295230
12. Adamichou C, Genitsaridi I, Nikolopoulos D, Nikoloudaki M, Repa A, Bortoluzzi A et al. Lupus or not? SLE Risk Probability Index (SLERPI): a simple, clinician-friendly machine learning-based model to assist the diagnosis of systemic lupus erythematosus. Ann Rheum Dis 2021; 80(6): 758-766. https://doi.org/10.1136/annrheumdis-2020-219069
13. Nalband S, Sundar A, Prince A, Agarwal A. Feature selection and classification methodology for the detection of knee-joint disorders. Comput Methods Programs Biomed 2016;127:94-104. https://doi.org/10.1016/j.cmpb.2016.01.020
14. Patro SGK, Sahu KK. Normalization: A Preprocessing Stage. arXiv [Pre-print] 2015. DOI: 10.48550/ARXIV.1503.06462

15. Stiglic G, Kocbek P, Fijacko N, Zitnik M, Verbert K, Cilar L. Interpretability of machine learning-based prediction models in healthcare. Wiley Interdiscip Rev Data Min Knowl Discov 2020;10(5). doi: https://doi.org/10.1002/widm.1379
16. Binder A, Ellis S. When to order an antinuclear antibody test. BMJ 2013; 347 :f5060 doi:10.1136/bmj.f5060
17. Wichainun R. Sensitivity and specificity of ANA and anti-dsDNA in the diagnosis of systemic lupus erythematosus: A comparison using control sera obtained from healthy individuals and patients with multiple medical problems. Asian Pac J Allergy Immunol Immunology 2013;31(4). doi: 10.12932/AP0272.31.4.2013
18. Arnaud L, Mathian A, Boddaert J, Amoura Z. Late-Onset Systemic Lupus Erythematosus. Drugs Aging 2012;29(3):181-189. doi: 10.3899/jrheum.080957
19. LALANI S, POPE J, de LEON F, PESCHKEN C. Clinical Features and Prognosis of Late-onset Systemic Lupus Erythematosus: Results from the 1000 Faces of Lupus Study. J Rheumatol 2009;37(1):38-44. doi: 10.3899/jrheum.080957
20. Beckwith H, Lightstone L. Rituximab in Systemic Lupus Erythematosus and Lupus Nephritis. Nephron Clin Pract 2014;128(3-4):250-254. doi: 10.1159/000368585
21. Mahajan A, Amelio J, Gairy K, Kaur G, Levy R, Roth D, et al. Systemic lupus erythematosus, lupus nephritis and end-stage renal disease: a pragmatic review mapping disease severity and progression. Lupus 2020;29(9):1011-1020. doi: 10.1177/0961203320932219
22. Yu H, Nagafuchi Y, Fujio K. Clinical and Immunological Biomarkers for Systemic Lupus Erythematosus. Biomolecules 2021;11(7):928. doi: 10.3390/biom11070928
23. Giannouli S. Anaemia in systemic lupus erythematosus: from pathophysiology to clinical assessment. Ann Rheum Dis 2006;65(2):144-148. doi: 10.1136/ard.2005.041673
24. Al Arfaj A, Khalil N. Clinical and immunological manifestations in 624 SLE patients in Saudi Arabia. Lupus (2009); 18(5):465–473. https://doi.org/10.1177/0961203308100660
25. AlSaleh J, Jassim V, ElSayed M, Saleh N, Harb D. Clinical and immunological manifestations in 151 SLE patients living in Dubai. Lupus 2008; 17(1):62-66. doi:10.1177/0961203307084297
26. El Hadidi K, Medhat B, Abdel Baki N, Abdel Kafy H, Abdelrahaman W, Yousri A, et al. Characteristics of systemic lupus erythematosus in a sample of the Egyptian population: a

retrospective cohort of 1109 patients from a single center. Lupus. 2018;27(6):1030-1038. doi:10.1177/0961203317751856

27. Al Rasbi A, Abdalla E, Sultan R, Abdullah N, Al Kaabi J, Al-Zakwani I, et al. Spectrum of systemic lupus erythematosus in Oman: from childhood to adulthood. Rheumatology international 2018; 38(9):1691–1698. https://doi.org/10.1007/s00296-018-4032-2
28. Al-Shujairi A, Elbadawi F, Al-Saleh J, Hamouda M, Vasylyev A, Khamashta M. Literature review of lupus nephritis From the Arabian Gulf region. Lupus 2022; 0(0). doi:10.1177/09612033221137248
29. Hancock J, Khoshgoftaar T. CatBoost for big data: an interdisciplinary review. J Big Data 2020;7(1). doi: 10.1186/s40537-020-00369-8
30. Prokhorenkova L, Gusev G, Vorobev A, Dorogush A, Gulin A. CatBoost: unbiased boosting with categorical features. arXiv [Pre-print] 2017 Version 5 doi: https://doi.org/10.48550/arXiv.1706.09516
31. Anghel A, Papandreou N, Parnell T, De Palma A, Pozidis H. Benchmarking and optimization of gradient boosting decision tree algorithms. arXiv preprint 2018. Version 3 doi: https://doi.org/10.48550/arXiv.1809.04559
32. Branco P, Torgo L, Ribeiro R. A Survey of Predictive Modeling on Imbalanced Domains. ACM Comput Surv 2016;49(2):1-50. doi: https://dl.acm.org/doi/10.1145/2907070
33. Pargent, F., Pfisterer, F., Thomas, J., Bischl, B. Regularized target encoding outperforms traditional methods in supervised machine learning with high cardinality features. Comput Stat 2022. https://doi.org/10.1007/s00180-022-01207-6

Table 1: Dataset's features and its occurrence

Feature Name	Categories	Occurrence in SLE Population (No. %, N=138)	Occurrence in Control Population (No. %, N=81)
Fever	Yes No	41 (29.7%) 97 (70.2%)	7 (8.6%) 74 (91.3%)
Acute cutaneous lupus (ACL)	Yes (Rash) No	63 (45.6%) 75 (54.3%)	7 (8.6%) 74 (91.3%)
Chronic cutaneous lupus	Yes No	5 (3.6%) 133 (96.3%)	0 81 (100%)
Oral ulcers	Yes	29 (20%)	0

	No	109 (79%)	81 (100%)
Alopecia	Yes No	57 (41.3%) 81 (58.7%)	4 (4.9%) 77 (95%)
Joint Involvement	Yes No	121 (87.7%) 17 (12.3%)	0 81 (100%)
Serositis	Yes No	9 (6.5%) 129 (93.5%)	0 81 (100%)
Renal disorders	Yes No	62 (44.9%) 76 (55%)	0 81 (100%)
Lupus Nephritis class	None (No Kidney biopsy) Class II Class III Class IV Class V	35 (25.3 %) 1 (0.4%) 4 (1.8%) 16 (7.3%) 5 (2 %)	0 0 0 0 0
Proteinuria	Yes No	51 (37%) 87 (63%)	0 81 (100%)
vasculitis	Yes No	12 (8.7%) 126 (91.3%)	0 81 (100%)
Neurologic Disorder	None Psychosis Seizure	121 (87.7%) 5 (3.6 %) 12 (8.7%)	81 (100%) 0 0
Hemolytic Anemia	Yes No	47 (34%) 91 (66%)	6 (7.4%) 75 (92.6%)
Leukopenia	Yes No	18 (13%) 120 (86.9%)	1 (1.2%) 80 (98.7%)
Thrombocytopenia	Yes No	11 (8%) 127 (92%)	0 81 (100%)
Anti-dsDNA	Positive Negative	102 (73.9%) 36 (26%)	2 (2.4%) 79 (97.5%)
Anti-Smith (Sm) antibody	Positive Negative	17 (12.3%) 121 (87.7%)	0 81 (100%)
Antiphospholipid Antibodies	Positive Negative	46 (33.3%) 92 (66.6%)	2 (2.5%) 79 (97.5%)
C3 Complement	Positive Negative	95 (68.8%) 43 (31.1%)	2 (2.5%) 79 (97.5%)
C4 Complement	Positive Negative	95 (68.8%) 43 (31.1%)	2 (2.5%) 79 (97.5%)
Rheumatoid factor	Positive Negative	18 (13%) 120 (86.9%)	0 81 (100%)
Gender	Male Female	5 (3.6%) 133 (96.4%)	12 (14.8) 69 (85.2%)
	20 years or less	16 (11.6%)	1 (1.2%)
	21 – 25 year	15 (10.8%)	5 (6.2%)
	26 – 30 year	25 (18.1%)	9 (11.1%)

Age Period	31 – 35 year	29 (21%)	11 (13.6%)
	36 – 40 year	16 (11.6%)	8 (9.9%)
	41 – 45 year	23 (16.6%)	7 (8.6%)
	46 – 50 year	5 (3.6%)	7 (8.6%)
	more than 50 year	6 (4.3%)	33 (40.7%)

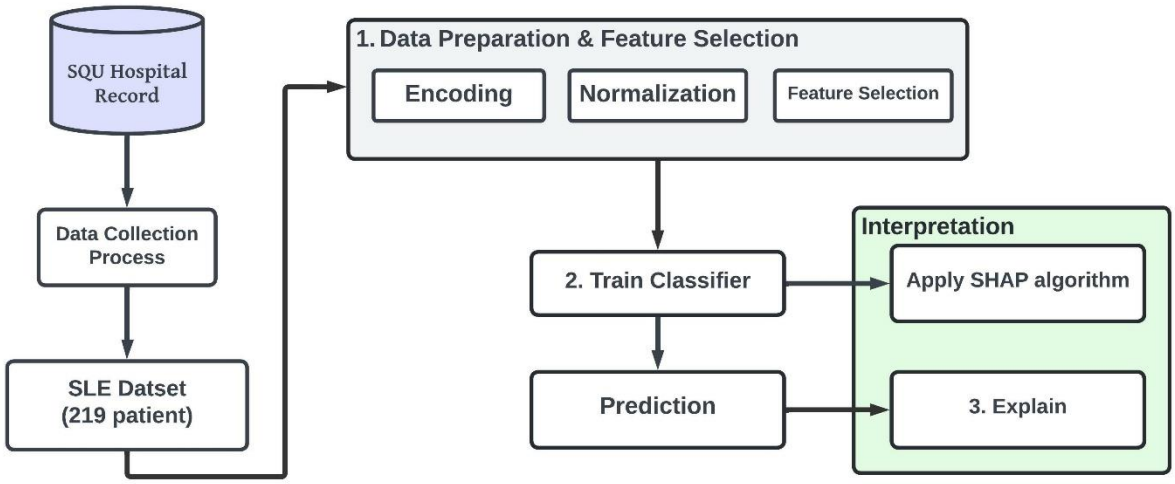
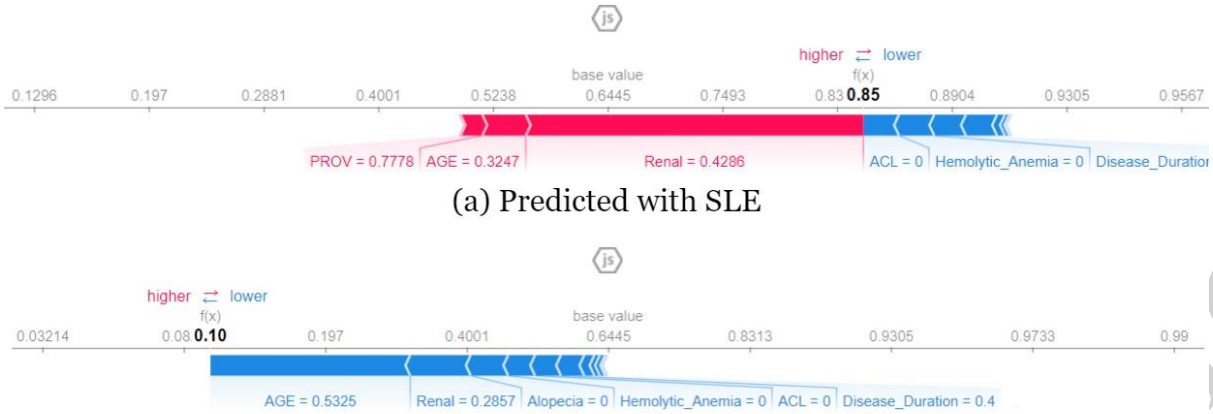


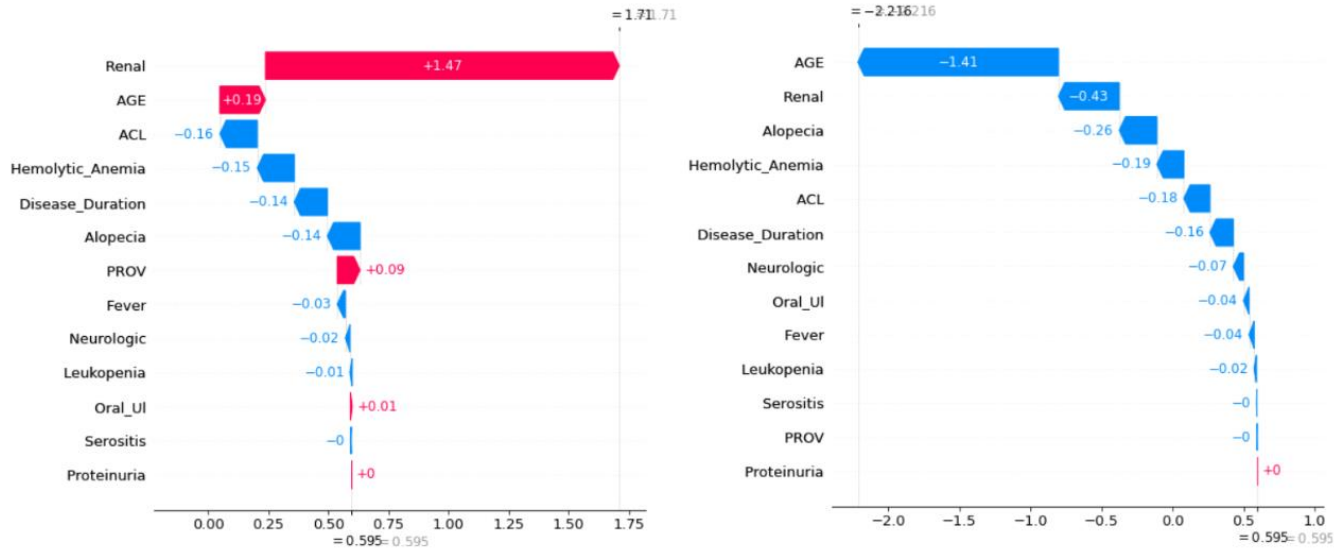
Figure 1: Flowchart of the three-stage interpretable framework.



(a) Predicted with SLE

(b) Predicted with No-SLE

Figure 2: force plot of CatBoost model prediction (values are normalized). f(x) is the predicted probability. The arrows in each plot show the direction of influence each predictor has over the payout i.e. the prediction. The colors are used to indicate the influence of the predictors, whether it increases (red) or reduces (blue) the probability of having SLE.



(a) Patient 1 (Predicted with SLE)

(b) Patient 2 (Predicted with No SLE)

Figure 3: waterfall plot of CatBoost model. The waterfall plot displays SHAP values representing feature contribution toward a positive prediction. It reflects the magnitude of influence each predictor had. The colors represent negative SHAP values for Blue, and positive SHAP values for Red.

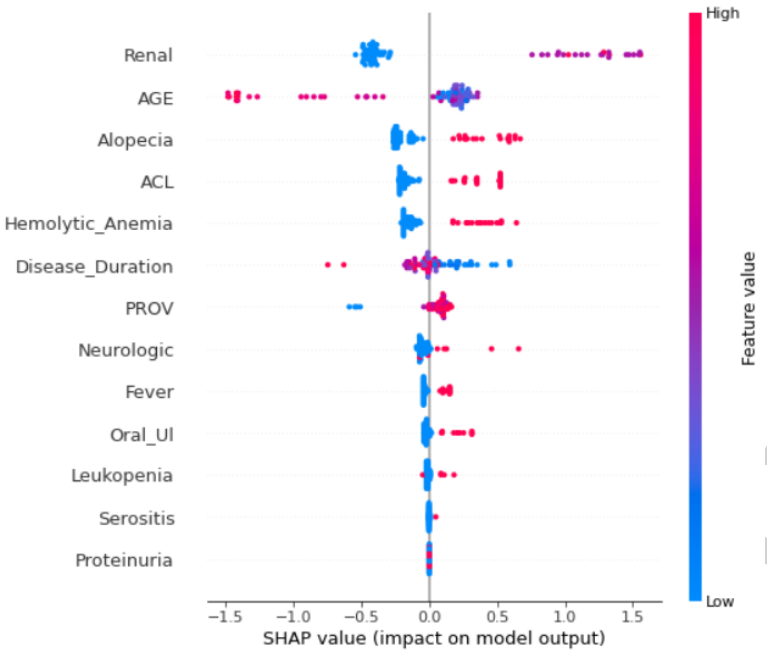


Figure 4: Summary plot of CatBoost model. The summary plot combines feature importance with feature effects. Each point on the summary plot is a Shapley value for a feature and an instance. The position on the y-axis is determined by the feature's importance and on the x-axis by the Shapley value. The summary plot is similar to the waterfall plot in ranking the contribution of all features based on SHAP values. The major difference between the two is that it is applied to the entire testing set rather than one single data observation. Each dot represents an observation from the testing set, and the color of the dot reflects the value of the associated feature. For example, in the feature 'AGE' red dots correspond to patients with high value i.e. old patients and Blue corresponds to young patients.

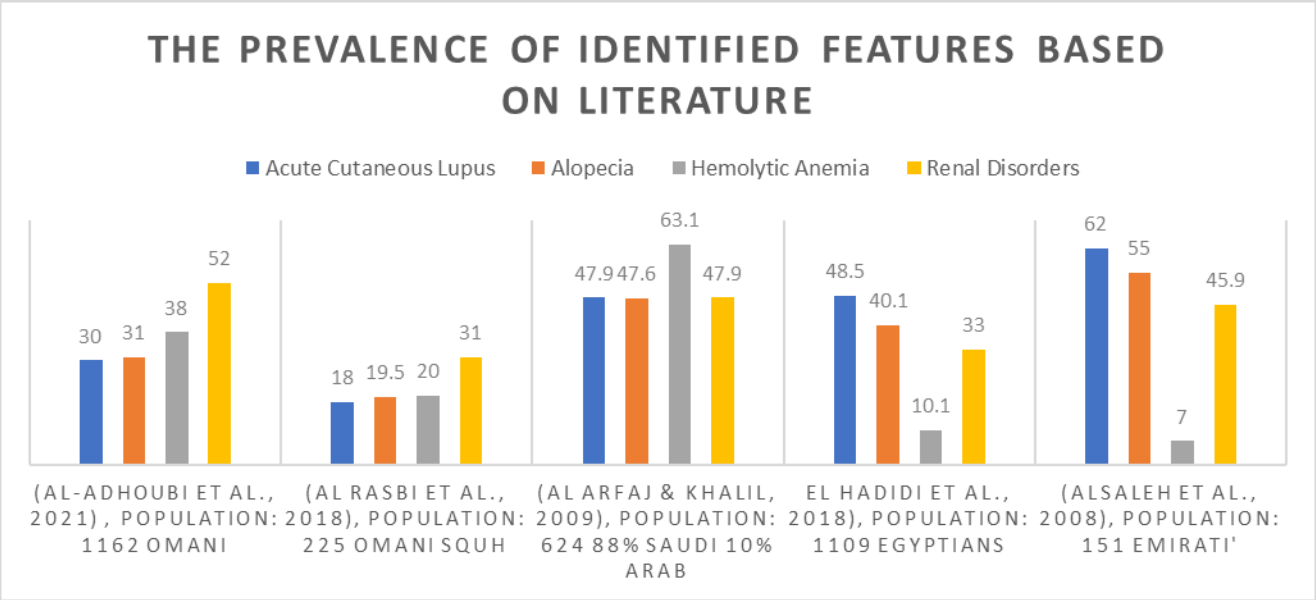


Figure 5: The frequency of the most influential features as shown by SHAP in cohorts across the Arab region.