Sultan Qaboos University Med J, August 2023, Vol. 23, Iss. 3, pp. 328–335, Epub. 28 Aug 23
Submitted 23 Jun 22
Revisions Req. 18 Sep & 14 Nov 22; Revisions Recd. 23 Oct & 24 Nov 22
Accepted 8 Dec 22
https://doi.org/10.18295/squmj.12.2022.069

CLINICAL & BASIC RESEARCH

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

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

abstract: ***Objectives:*** This study aimed to 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:*** Data of 219 patients from 2006 to 2019 were extracted from Sultan Qaboos University Hospital's electronic records. Among these, 138 patients had SLE, while the remaining 81 had other rheumatologic diseases. Clinical and demographic features were analysed to focus on the early stages of the disease. Recursive feature selection was implemented to choose the most informative features. The CatBoost classification algorithm was utilised to predict SLE, and the SHAP explainer algorithm was applied on top of the CatBoost model to provide individual prediction reasoning, which was then validated by rheumatologists. ***Results:*** CatBoost achieved an area under the receiver operating characteristic curve score of 0.95 and a sensitivity of 92%. The SHAP algorithm identified four clinical features (alopecia, renal disorders, acute cutaneous lupus and haemolytic anaemia) and the patient's age as having the greatest contribution to the prediction. ***Conclusion:*** An explainable framework to predict SLE in patients and provide reasoning for its prediction was designed and validated. This framework enables clinicians to implement early interventions that will lead to positive healthcare outcomes.

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

Advances in Knowledge

- *The first self-explainable prediction framework for systemic lupus erythematosus (SLE) specific to the Omani population was developed.*
- *The framework achieved an area under the receiver operating characteristic curve score of 0.956 and had a sensitivity of 92%.*
- *It identified patterns in clinical manifestation that are unique to the Omani population.*
- *Four clinical features (alopecia, renal disorders, acute cutaneous lupus and haemolytic anaemia) had the highest contribution to the model's prediction.*
- *Compared to other Arab ethnicities, the frequency of renal disorders in Oman was the highest, while the frequency of alopecia 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, thus prompting further investigation until an official diagnosis is made.*
- *It will enable clinicians to compare the information reported by the model with their own knowledge through an interpretation algorithm, thereby increasing the probability of correct diagnosis and encouraging the adoption of machine learning (ML) in healthcare.*
- *It provides a practical introduction of ML and interpretation tools to the medical diagnosing process that improves early detection of SLE; a crucial factor in lowering flare rates and reducing mortality.*

Systemic lupus erythematosus (SLE) is a chronic multisystem autoimmune disease caused by genetic and environmental factors that contribute to the production of high-titre autoantibodies targeting native DNA and other cellular elements.[1] The creation of these autoantibodies leads to a pathological process that manifests in different medical conditions in various organ systems, ranging from skin arthralgia to cardiovascular and renal morbidity.[2] The clinical phenotype of SLE varies with race, gender and age, which makes the disease challenging to diagnose.[3] In Oman, the mortality rate of SLE is estimated to be 5%, with a mean prevalence of 38 per 100,000; this is higher than the SLE prevalence in Saudi Arabia and lower than the SLE prevalence in the UAE.[4] Initial SLE symptoms are often non-specific and mimic other medical conditions, thus increasing the risk 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, significant improvements have been made in treatment strategies for SLE. However,

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*

despite the improved prognosis, various challenges remain for the diagnosis and therapeutic management of SLE.[5] One of those challenges is early diagnosis. The onset of SLE is gradual, and clinically evident manifestations develop over the years. Moreover, a variety of conditions may mimic SLE, including infectious and haematologic diseases.[6] It has been proven through database analysis that patients with a diagnosis window of less than six months (between probable SLE onset and diagnosis) had lower flare rates and hospitalisations compared to patients with a late diagnosis.[7] Late diagnosis is also associated with the risk of developing progressive organ damage, which subsequently increases the mortality rate.[8]

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

Methods

The data set used in this study was collected from structured and unstructured sources. This included the TrakCare electronic medical records in Sultan Qaboos University Hospital's Rheumatology Clinic, Muscat, Oman. TrakCare stores patients' information, medical states and medical histories. Patients' demographic data were obtained directly from TrakCare. Meanwhile, clinical data were unstructured as they were stored in the patients' medical histories as clinical notes from each visit to the hospital. The inclusion criteria for rheumatology patients were a positive antinuclear antibody (ANA) test, while the exclusion criteria included non-Omani patients as well as those with insufficient data. To separate patients with SLE and control diseases, the most recent SLE classification criteria set by EULAR/ACR were used, wherein patients with a score of 10 or above are diagnosed with SLE.[9] A total of 219 patient records matched the inclusion criteria: 138 were diagnosed with SLE and 81 had other control diseases; this was also validated by a rheumatologist on a case-by-case basis.

The framework consists of three main stages: (1) feature selection, which reduces noisy data and utilises only the most informative features; (2) classification, in which the classifier trains and tests the model to predict the presence of SLE; and (3) prediction, during which the explainer algorithm provides individual prediction breakdowns through informative visual plots after the model has been trained.

For the first stage, the recursive feature elimination (RFE) algorithm with 10-fold cross-validation (CV) was used [Figure 1]. The algorithm works by building a model, selecting the best feature, discarding the selected feature and then repeating this process for the remaining features until all features have been traversed.

For the second stage, Categorical Boosting (or CatBoost), an ensemble learning algorithm that is based on gradient boosting was implemented.[10]

For the final stage, the SHAP library was implemented.[11] The SHAP method 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 magnitude of the Shapley value represents the importance of the feature relative to the prediction. The SHAP tool can also perform local and global interpretability simultaneously. With the help of the SHAP algorithm, each prediction can be broken down individually. As a demonstration, two individuals were selected from the testing set: one predicted to have the disease and one predicted not to have it. 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 contribute to the model's prediction for a specific observation. The colours in the force plot correspond to the feature pushing the prediction probability higher or lower. The model's target 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, a waterfall plot is used. 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 was used to train the model and the latter was used to test the model's performance. Additionally, a subset of the training data set was used for CV to protect the models from overfitting and optimise the model's parameters. Each of the models

Table 1: Features and occurrence of SLE among the included study sample (N = 219)

Feature	n (%)	
	Occurrence in SLE population (n = 138)	Occurrence in control population (n = 81)
Gender		
Male	5 (3.6)	12 (14.8)
Female	133 (96.4)	69 (85.2)
Age in years		
≤20	16 (11.6)	1 (1.2)
21–25	15 (10.8)	5 (6.2)
26–30	25 (18.1)	9 (11.1)
31–35	29 (21.0)	11 (13.6)
36–40	16 (11.6)	8 (9.9)
41–45	23 (16.6)	7 (8.6)
46–50	5 (3.6)	7 (8.6)
>50	6 (4.3)	33 (40.7)
Fever		
Yes	41 (29.7)	7 (8.6)
No	97 (70.2)	74 (91.3)
Acute cutaneous lupus		
Yes (Rash)	63 (45.6)	7 (8.6)
No	75 (54.3)	74 (91.3)
Chronic cutaneous lupus		
Yes	5 (3.6)	0
No	133 (96.3)	81 (100)
Oral ulcers		
Yes	29 (20.0)	0
No	109 (79.0)	81 (100)
Alopecia		
Yes	57 (41.3)	4 (4.9)
No	81 (58.7)	77 (95.0)
Joint involvement		
Yes	121 (87.7)	0
No	17 (12.3)	81 (100)
Serositis		
Yes	9 (6.5)	0
No	129 (93.5)	81 (100)
Renal disorders		
Yes	62 (44.9)	0
No	76 (55.1)	81 (100)
Lupus nephritis class		
None (no kidney biopsy)	35 (25.3)	0
Class II	1 (0.4)	0
Class III	4 (1.8)	0
Class IV	16 (7.3)	0
Class V	5 (2.0)	0
Proteinuria		
Yes	51 (37.0)	0
No	87 (63.0)	81 (100)
Vasculitis		
Yes	12 (8.7)	0
No	126 (91.3)	81 (100)
Neurologic disorder		
None	121 (87.7)	81 (100)
Psychosis	5 (3.6)	0
Seizure	12 (8.7)	0
Haemolytic anaemia		
Yes	47 (34.0)	6 (7.4)
No	91 (66.0)	75 (92.6)
Leukopaenia		
Yes	18 (13.0)	1 (1.2)
No	120 (86.9)	80 (98.7)
Thrombocytopaenia		
Yes	11 (8.0)	0
No	127 (92.0)	81 (100)
Anti-dsDNA		
Positive	102 (73.9)	2 (2.4)
Negative	36 (26.0)	79 (97.5)
Anti-Smith antibody		
Positive	17 (12.3)	0
Negative	121 (87.7)	81 (100)
Antiphospholipid antibodies		
Positive	46 (33.3)	2 (2.5)
Negative	92 (66.6)	79 (97.5)
C3 complement		
Positive	95 (68.8)	2 (2.5)
Negative	43 (31.1)	79 (97.5)
C4 complement		
Positive	95 (68.8)	2 (2.5)
Negative	43 (31.1)	79 (97.5)
Rheumatoid factor		
Positive	18 (13.0)	0
Negative	120 (86.9)	81 (100)

SLE = systemic lupus erythematosus.

underwent a hyper-parameter optimisation through grid search with a five-fold CV. To avoid reporting biased results and limit overfitting, the measurement's average of 10 repetitions was calculated for each model. Finally, three other classifiers were evaluated similarly: the multi-layer perceptron, support vector machine (SVM) and random forest. Their performance evaluations were compared to CatBoost to observe the effectiveness of CatBoost. The classifiers were selected

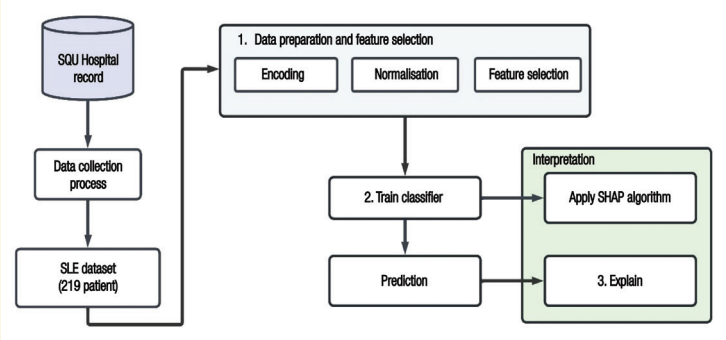

Figure 1: Flowchart of the three-stage interpretable framework of the study.
SQU = Sultan Qaboos University; SLE = systemic lupus erythematosus; SHAP = SHapley Additive exPlanations.

based on related studies that employed ML for disease prediction.[12,13]

Due to the imbalanced nature of the problem, the area under the ROC curve (AUC) and sensitivity parameters were 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 (MREC: #1418 and #1650). No participant consent was required for this study, as per the regulation of Sultan Qaboos University Hospital.

Results

The extracted data encompasses patient records from January 2006 to December 2019. The majority of the records represented females (92%). Patients between 25 years old and their late 30s represented the largest age group, with a mean age of 38 years. The had the highest number of patients (37.9%) followed by Muscat (23.7%).

The initial data contained 28 clinical, demographic and laboratory variables ('features' in ML), and no missing values were found in the data [Table 1]. The laboratory features included results from immunological tests such as the anti-dsDNA test and anti-Smith antibody, among others. These features, however, are highly sensitive to SLE and can introduce bias to the prediction model; therefore, they were dropped. The remaining data consisted of 20 clinical and demographic features. The majority of the features were represented by non-numerical (categorical) values. This entailed the transformation (encoding) to numerical values, a prerequisite for all statistical models. Thus, ordinal encoding was applied. Furthermore, due to the variance in range for different features, min-max normalisation was also applied.[14]

After applying the RFE algorithm, the optimal number of features selected was 13. From the RFE-selected features, three demographic features and 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, respectively. CatBoost had 92% sensitivity, Random Forest had 89% and SVM had 86%.

Two samples from the testing set were used to generate the different SHAP plots. The first sample (patient one) was predicted to have SLE; the force plot attributed this to renal disorders and the patient's age [Figure 2A]. Since the values were normalised, they were cross-referenced with the test data and it was found that the patient's age was 40 years, which

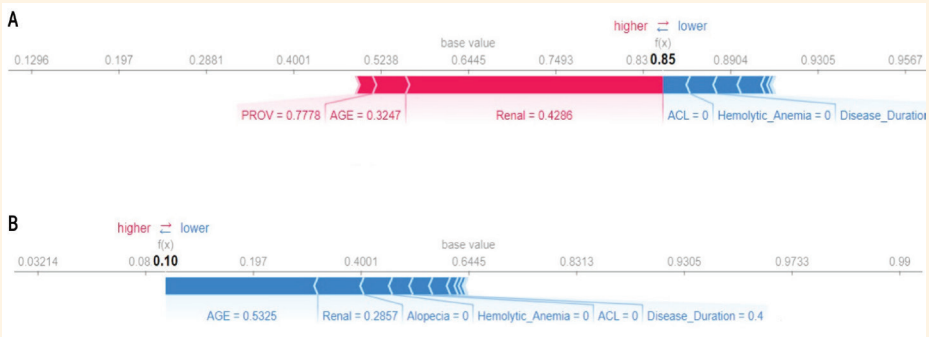

Figure 2: Force plot of CatBoost model prediction (values are normalised). F(x) is the predicted probability. The arrows in each plot show the direction of influence each predictor has over the pay-out (i.e. the prediction). The colours are used to indicate the influence of the predictors: whether it increases (red) or decreases (blue) the probability of having systemic lupus erythematosus.

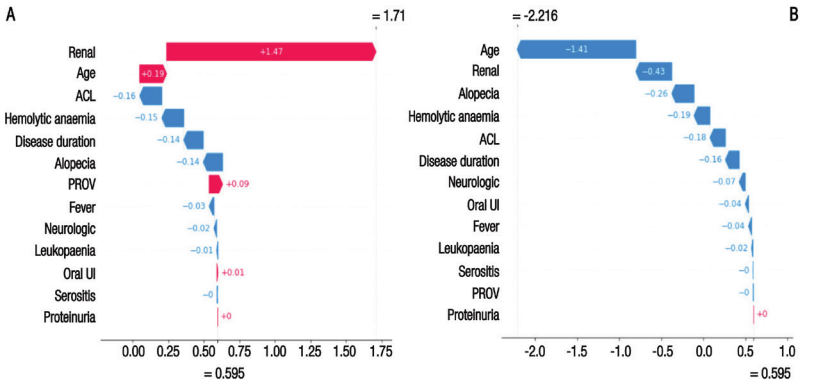

Figure 3: Waterfall plot of CatBoost model. The waterfall plot displays SHapley Additive exPlanations (SHAP) values that represent feature contribution towards a positive prediction. It reflects the magnitude of influence each predictor had. The colours represent negative SHAP values (blue) and positive SHAP values (red).

ACL = acute cutaneous lupus; Ul = ulcers.

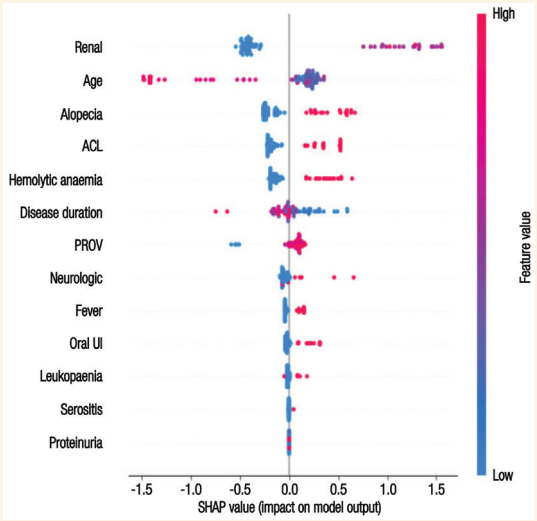

Figure 4: Waterfall plot of CatBoost model. The waterfall plot displays SHapley Additive exPlanations (SHAP) values that represent feature contribution towards a positive prediction. It reflects the magnitude of influence each predictor had. The colours represent negative SHAP values (blue) and positive SHAP values (red).

ACL = acute cutaneous lupus; Ul = ulcers.

falls within the age group in which SLE is most active. Additionally, the patient had been diagnosed with lupus nephritis (LN), a disease that is commonly caused by an auto-immune disorder. On the other hand, the second sample (patient two) displayed a lack of any autoimmune manifestation and long disease duration [Figure 2B]. Furthermore, the patient's age of 56 years placed him outside the age group in which SLE is most active.

Looking at the waterfall plot for patient one, the feature with the highest SHAP value is a renal disorder by a large margin [Figure 3A]. Due to its high SHAP value, the presence of renal disorder in Patient 1 contributed most 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 towards class 0. The non-existence of alopecia, acute cutaneous lupus (ACL) and haemolytic anaemia in patient one's 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 two indicates that age is the greatest contributor towards class 0, followed by the absence of any renal disorders [Figure 3B].

Similar to what was deduced, it can be seen that the older the patient was, the less likely they are to have SLE, as evidenced by the red dots on the negative scale of the SHAP values [Figure 4]. The same can

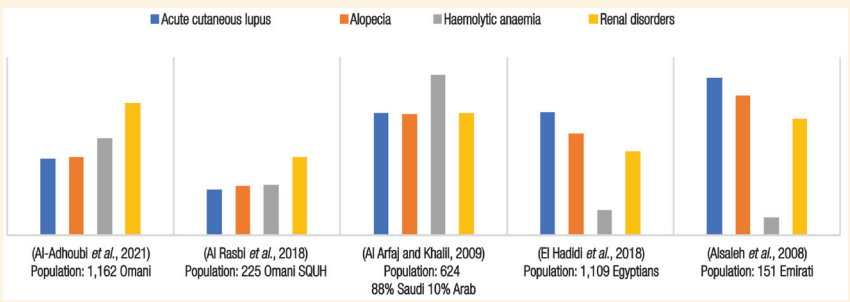

Figure 5: The frequency of the most influential features as shown by SHAP in cohorts across the Arab region.

be said for disease duration: it was found that long disease durations without autoimmune manifestation correlated with the absence of SLE. This 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 contributing features to SLE, followed by alopecia, ACL and haemolytic anaemia. The lowest contributing features are serositis, proteinuria and leukopaenia.

Discussion

In clinical applications, the ability to justify a prediction is as important as the prediction score itself. This is due to the high sensitivity of the medical environment, where misclassification could lead to devastating results. Therefore, trusting complex ML models is challenging 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 labelled affects the model's results.[15] As a result, interpretable ML has emerged as an area of research aiming to design transparent and explainable models by developing the means to transform black-box ML models into white-box ML models. By providing transparent predictions, domain experts can accurately interpret the results meaningfully.

Through the use of the SHAP algorithm, clinicians can understand the model's reasoning, thus resulting in it resembling clinical reasoning. The model designed in this study is intended for use between early to mid-screening, suggesting implementation when 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 and anti-dsDNA tests are highly sensitive and decisive if found positive.[17] Additionally, an immunologist compared multiple individual prediction breakdown plots and validated the results and the model justification.

Some of the features used to profile the patients were age, age-onset and disease duration. 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 anaemia or LN, were statistically shown to be less likely to be diagnosed with SLE. Experts have pointed out, however, that the intensity of SLE increases and decreases at intervals differently from patient to patient; therefore, on 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]

According to SHAP, renal disorders were the highest contributing feature. This concurred with the findings of Beckwith and Lightstone, who stated that approximately 40–70% of SLE patients develop clinically diagnosed renal involvement, known as LN.[20] Lupus nephritis is commonly diagnosed through a 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 *et al.*, who found that 52% of SLE patients developed LN.[4] Despite the majority of this study's data lacking kidney biopsy information, LN was also present in 11% of the patients with renal disorders.

Moreover, other clinical features that had about the same influence on the prediction were found. These were alopecia, ACL and haemolytic anaemia. Alopecia is 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. Acute cutaneous lupus, which includes a butterfly rash across the face between the eyes and nose, is a sign of VGLL-3 and anti-SSA antibodies, which indicate skin damage activity caused by lupus.[22] Haemolytic anaemia is the most common blood disorder, affecting about half of all people with active lupus.[23] Anaemia is caused by a shortage of healthy red blood cells needed by the body to carry oxygen to the body's tissues. Haemolytic anaemia, 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, the UAE and Egypt [Figure 5].[24–26] Acute cutaneous lupus or skin rash was found to be more prevalent in all other Arab cohorts, reaching as high as 62% in the UAE. Haemolytic anaemia was the most varying feature in Egypt and the UAE but is less prevalent than in Oman, while in Saudi Arabia, it is more prevalent than in Oman.[27] Renal disorders remained high, with approximately 50% of all the cohorts having some renal damage except for a slight decrease to 33% in Egypt.

Studies also indicated that out of all renal biopsies, approximately 10–36% of patients are diagnosed with LN in the Gulf region. Lupus nephritis also tends to run a severe course in Gulf populations with a high incidence of Class IV LN.[28]

Overall, with three out of four critical features found to be more prevalent in other Arab ethnicities, this study's 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 are presented by autoimmune diseases in general. However, classification models can be trained to detect patterns specific to the Omani population. These patterns are the basis for the model's prediction of SLE presence.

These findings help identify patterns in clinical manifestations that are unique to the Omani population and the Arab region by employing explainable prediction. Moreover, this research also highlights the CatBoost algorithm, which has received widespread attention in recent years for its fast calculation speed, powerful generalisation ability and strong predictive performance.[29–31] A improvement margin of 0.21 AUC over the other classifiers was achieved; this may be attributed to Cat-Boost's novel implementation of ordered boosting and permutation-driven alternative to the classic algorithm. This study also acknowledges the problem with imbalance classification evaluation, where the research is biased towards the performance of cases that are poorly represented in the data samples.[32] Standard evaluation criteria tend to focus on the most frequent cases, which could lead to sub-optimal classification models if applied. Therefore, AUC and sensitivity were selected as the evaluation criteria for this study.

Finally, by combining the framework's prediction with the interpretation algorithm, self-explainable frameworks that enable physicians to make meaningful decisions based on ML-based information combined with their knowledge are being promoted. This improves the probability of correct diagnosis and encourages the adoption of ML in healthcare. However, the retrospective nature of the data hinders the achievement of these goals. An ideal framework would be much more effective with longitudinal data of SLE patients that include pre-diagnosis profiles before the appearance of adverse symptoms. Moreover, this study's framework may not scale properly with large data sets. 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 the model, for example, the outcome can be presented as a prediction probability instead of a binary value.

Conclusion

This study proposed a three-stage interpretable framework for predicting the presence or absence of SLE in an Omani cohort of 219 patients. A CatBoost classifier and SHAP interpretation tool were implemented to predict and justify individual predictions and eliminate any risk of misclassification. In addition to the patient's age, four clinical features were identified to have the highest influence on the prediction: alopecia, renal disorders, ACL and haemolytic anaemia. These features serve as indicators of lupus activity at varying rates. When combined with the patient's age and age-onset, the model was able to establish a profile of the disease relative to Omanis. Overall, these findings aid in providing a practical introduction of ML and interpretation tools to medical diagnosis, thereby increasing the efficiency of medical testing and subsequently enabling early intervention, which can lead to better treatment and positive healthcare outcomes.

AUTHORS' CONTRIBUTION

HZ and AA conceived the idea and designed the study. SA collected the data, while BA and AA-A 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.

ACKNOWLEDGEMENT

The authors would like to thank The Research Council in Oman for supporting this research. They also gratefully acknowledge the support from and cooperation of the Sultan Qaboos University Hospital.

CONFLICT OF INTEREST

The authors declare no conflicts of interest.

FUNDING

This work was supported in part by an internal grant (IG/SCI/COMP/21/02).

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; 28:114–21. https://doi.org/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; 111:1298–300. https://doi.org/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. https://doi.org/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; 24:847–54. https://doi.org/10.1111/1756-185x.14130.

5. Felten R, Lipsker D, Sibilia J, Chasset F, Arnaud L. The history of lupus throughout the ages. J Am Acad Dermatol 2022; 87:1361–9. https://doi.org/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; 10:243. https://doi.org/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; 12:179–90. https://doi.org/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:e031850. https://doi.org/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; 80:775–81. https://doi.org/10.1136/annrheumdis-2020-219373.

10. Dorogush A, Ershov V, Gulin A. CatBoost: Gradient boosting with categorical features support. arXiv [Pre-print] 2018. https://doi.org/10.48550/arXiv.1810.11363.

11. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. 31st International Conference on Neural Information Processing Systems. Adv Neural Inf Process Syst 2017; 30:4768–77.

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; 80:758–66. 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. https://doi.org/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:e1379. https://doi.org/10.1002/widm.1379.

16. Binder A, Ellis S. When to order an antinuclear antibody test. BMJ 2013; 347:f5060. https://doi.org/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 2013; 31:292–8. https://doi.org/10.12932/AP0272.31.4.2013.

18. Arnaud L, Mathian A, Boddaert J, Amoura Z. Late-onset systemic lupus erythematosus. Drugs Aging 2012; 29:181–9. https://doi.org/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; 37:38–44. https://doi.org/10.3899/jrheum.080957.

20. Beckwith H, Lightstone L. Rituximab in systemic lupus erythematosus and lupus nephritis. Nephron Clin Pract 2014; 128:250–4. https://doi.org/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; 29:1011–20. https://doi.org/10.1177/0961203320932219.

22. Yu H, Nagafuchi Y, Fujio K. Clinical and immunological biomarkers for systemic lupus erythematosus. Biomolecules 2021; 11:928. https://doi.org/10.3390/biom11070928.

23. Giannouli S. Anaemia in systemic lupus erythematosus: From pathophysiology to clinical assessment. Ann Rheum Dis 2006; 65:144–8. https://doi.org/10.1136/ard.2005.041673.

24. Al Arfaj A, Khalil N. Clinical and immunological manifestations in 624 SLE patients in Saudi Arabia. Lupus 2009; 18:465–73. 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; 17:62–6. https://doi.org/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; 27:1030–8. https://doi.org/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. Rheumatol Int 2018; 38:1691–8. 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 2023; 32:155–65. https://doi.org/10.1177/09612033221137248.

29. Hancock J, Khoshgoftaar T. CatBoost for big data: An interdisciplinary review. J Big Data 2020; 7:94. https://doi.org/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; 5. 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 [Pre-print] 2018; 3. 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; 49:1–50. https://doi.org/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; 37:2671–92. https://doi.org/10.1007/s00180-022-01207-6.