CatBoost model with self-explanatory capabilities for predicting SLE in OMAN population

Hamza Zidoum[1*], Ali AlShareedah[1], Aliya Al-Ansari[2], Batool Al Lawati[3], S. Al-Sawafi[1]

[1] *Department of Computer, Science Sultan Qaboos, University, Muscat (Oman)*
[2] *Department of Biology, Sultan Qaboos, University, Muscat (Oman)*
[3] *Department of Medicine Sultan Qaboos University, Muscat (Oman)*

Abstract

Systemic lupus erythematosus (SLE) presents as an autoimmune condition influenced by both genetic and environmental factors, showcasing a diverse range of clinical symptoms and often, unpredictable disease flares. Despite advancements in classification methods, the timely diagnosis of SLE remains a challenge for many patients. This research introduces an interpretable disease classification model that combines the robust predictive capabilities of CatBoost with the transparent interpretation tools offered by SHapley Additive exPlanations (SHAP). Trained on a local cohort comprising 219 Omani patients diagnosed with SLE and individuals with other control diseases, the CatBoost model demonstrates high performance. Moreover, utilizing the SHAP library enables the generation of individualized explanations for the model's decisions, highlighting key clinical features such as alopecia, renal disorders, cutaneous lupus, and hemolytic anemia, alongside patient age, which significantly contribute to the prediction process. The model achieved notable metrics, including an AUC score of 0.945 and an F1-score of 0.92, underscoring its efficacy in SLE prediction

Keywords

Systemic lupus erythematosus; CatBoost; Feature selection; Model interpretation; SHAP

1. Introduction

Systemic lupus erythematosus (SLE) stands as a chronic autoimmune disease affecting multiple systems of the body. Its clinical presentation varies across different races, genders, and age groups, rendering diagnosis challenging [1]. Despite significant strides in SLE treatment strategies, diagnostic and therapeutic hurdles persist [2]. Early diagnosis remains particularly problematic, given the gradual onset of SLE symptoms over years, alongside the potential for other conditions to mimic its manifestations, including infectious and hematologic diseases [3]. Data analysis underscores the importance of diagnosing SLE within a narrow window, as delayed diagnosis correlates with increased flare rates, hospitalizations, and the risk of progressive organ damage, ultimately elevating mortality rates [4]. In Oman, where the mortality rate stands at 5% and the mean prevalence at 38 per 100,000 individuals, limited research exists on the specific clinical and serologic characteristics of the Omani population. [5], [6]. This study contributes significantly on three fronts: firstly, by identifying unique clinical manifestation patterns specific to the Omani population, filling a notable gap in the literature. Secondly, by introducing the CatBoost algorithm, renowned for its rapid computation, strong generalization, and high predictive accuracy, alongside leveraging advanced machine learning techniques such as the SHAP algorithm, RFECV-based feature selection, and GridSearchCV-based hyperparameter optimization. Thirdly, by integrating the model's predictions with interpretability

Late-breaking work, Demos and Doctoral Consortium, colocated with The 2nd World Conference on eXplainable Artificial Intelligence: July 17–19, 2024, Valletta, Malta

*Correspondint author
zidoum@squ.edu.com (H. Zidoum), alansari@squ.edu.om (A. Al-Ansari), sksawafi@squ.edu.om (S. Al-Sawafi)
https://orcid.org/0000-0003-0365-650X (H. Zidoum)

algorithms, this research promotes self-explanatory models that empower physicians to cross-reference model outputs with their expertise, enhancing diagnostic accuracy and fostering the adoption of machine learning in healthcare.

2. Method

DATASET

The dataset utilized in this study originates from the Rheumatology clinic at Sultan Qaboos University Hospital. Approval for the study was granted by the Ethics Committee of the College of Medicine and Health Science at Sultan Qaboos University (MERC # 1418 and 1650). Data extraction involved both structured and unstructured sources, including the hospital's Electronic Medical Record (EMR) system named TrakCare, which stores patients' demographic information, medical states, and histories. While demographic data were directly obtained from TrakCare, clinical data were unstructured and retrieved from patients' medical histories in the form of clinical notes from each hospital visit. The dataset comprised records of Omani patients from 2006 to 2019 who met the entry criteria outlined by EULAR/ACR, which necessitate a positive Antinuclear Antibodies test (ANA test) followed by the application of additional classification criteria. Non-Omani patients and those with insufficient data were excluded. The dataset encompassed 214 Omani patient records, with 138 diagnosed with SLE, confirmed by rheumatologist assessment on a case-by-case basis, while the remaining 81 patients had other control diseases. Analysis revealed a female predominance of 92% and a mean age of 38. The Al Batinah Governorate accounted for the highest proportion of patients (37.9%), followed by Muscat (23.7%).

FEATURE SELECTION

Initial data comprised 20 clinical and demographic variables (referred to as "features" in machine learning), with no missing values detected. Categorical features were encoded using Ordinal encoding, and Min-Max normalization was applied due to variations in feature ranges. To enhance signal-to-noise ratio and select the most informative features, recursive feature elimination (RFE) based on Random Forest (RF) with ten-fold cross-validation (CV) was employed. RFE iteratively builds models, identifies the best feature, selects it, and repeats the process until all features are traversed.

Table 1
Characteristics of patients.

Feature	Category	Occurrence (No. %, N=219)
Fever	Yes/No	48 (21.9%)
Acute cutaneous lupus (ACL)	Yes/No	70 (31.9%)
Chronic cutaneous lupus	Yes/No	5 (2.28%)
Oral ulcers	Yes/No	29 (13.2%)
Alopecia	Yes/No	61 (27.8%)
Joint Involvement	Yes/No	202 (92.2%)
Serositis	Yes/No	9 (4%)
Renal Manifestation	Yes/No	62 (28.3%)
Lupus Nephritis class	None	112 (51 %)
	Class II	1 (0.4%)
	Class III	4 (1.8%)
	Class IV	16 (7.3%)
	Class V	5 (2 %)

Proteinuria	Yes/ No	51 (23%)
Vasculitis	Yes/ No	12 (5.4%)
Neurologic Disorder	None	121 (55.2 %)
	Psychosis	
	Seizure	5 (2.2 %)
Hemolytic Anemia	Yes/ No	12 (5.4%)
Leukopenia	Yes/ No	53 (24%)
Thrombocytopenia	Yes/ No	19 (8.6%)
		11 (5%)

We have developed four ML models to predict the presence or absence of SLE (Figure 1). In addition to three common ML models that are multi-layer perceptron (MLP), support vector machine (SVM), and Random Forest, we introduced CatBoost [11]

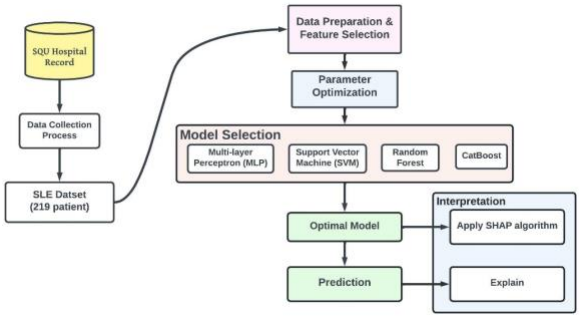


Figure 1: Flowchart of the development and evaluation process

CatBoost is an ensemble learning algorithm, similar to gradient boosting, but with some unique features. Its implementation of ordered boosting helps it handle categorical variables more effectively, which is particularly useful in real-world datasets where categorical features are common. The Oblivious Tree structure and random permutations are additional techniques that contribute to its robustness and efficiency especially when dealing with categorical data. Ordered boosting calculates leaf values during the selection of the tree structure to reduce overfitting. Oblivious Tree structure is used to construct CatBoosts' model ensembles which means that all the leaves are in the same level and the same splitting criterion is applied to all intermediate nodes within the same level of tree. The use of Oblivious Tree structure greatly improves the performance speed and efficiency. Random permutations of the training examples are also applied to fight the prediction shift caused by a special kind of target leakage present in all existing implementations of gradient boosting algorithms [12].

To train and validate the performance of our model, the dataset was divided into two parts in a 70:30 ratio (i.e. 70% of the dataset is used for training and 30% for testing). 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.

Due to the imbalanced nature of the data set, several parameters are used to evaluate the classification performance such as Recall, Specificity, F1-score, and AUC (Area Under ROC Curve). The problem with using imbalanced data set for classification is that the user is biased to the performance on cases that are poorly represented in the data samples [13]. 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. Each of the models undergoes a hyper-parameter optimization through grid search with a five-fold cross-validation. Finally, to avoid reporting biased results and limit overfitting, we calculated the average of 10 repetitions for each model.

MODEL INTERPRETATION

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 consequences. It is therefore challenging to trust complex ML models for a number of 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 [14]. For these reasons and more, Interpretable ML has thus emerged as an area of research that aims to design transparent and explainable models through developing means to transform a black-box ML model into a white-box ML model. By providing transparent prediction, domain experts can accurately interpret the results meaningfully. In 2017, Lundberg and Lee proposed a unified framework to interpret ML predictions. SHapley Additive exPlanations (SHAP) is derived from 'Shapley values', a concept that is commonly used within the field of cooperative game theory to determine the payout for each player within a cooperative coalition [14]. Casting this concept onto prediction models, the payout is mapped as the final prediction while players are mapped as the model's features. The contribution of a feature to the final prediction can be determined by looking at the magnitude and sign of the Shapley value. Specifically, the importance of a feature relative to the payout (prediction) is represented by the magnitude of the related Shapley value. More importantly, this framework provides local and global interpretability simultaneously

3. Results

Applying the RFE feature selection algorithm resulted in 13 optimal features (Figure 2). In Table 3, the features that were selected are indicated with a True value. Overall, three demographic features, as well as 10 clinical features, were selected

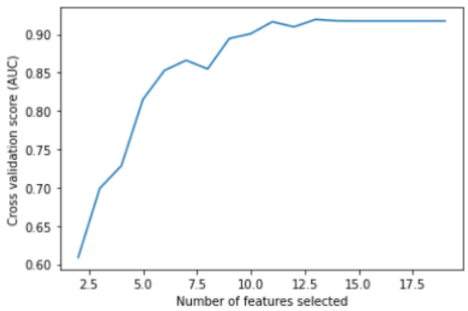


Figure 2: Visualizing RFE's optimal number of features with 10-fold CV

Table 2
Comparing feature sets obtained from different feature selection algorithms.

Feature Name	Selected by RFE
AGE	True
Disease Duration	True
PROV	True
Fever	True
ACL	True
Oral_Ul	True
Alopecia	True
Serositis	True
Renal	True
Proteinuria	True

Neurologic	True
Hemolytic_Anemia	True
Leukopenia	True
Age onset	False
Thrombocytopenia	False

Comparing between the performance of the different classifiers (Table 4), CatBoost had the highest AUC score of 0.956 providing a slight edge in performance (Figure 3). This superiority in performance was also indicated in benchmarks against other recent classifiers (e.g. XGBoost and LightGBM) on a set of popular publicly available datasets [12]. In training phase, each set of decision trees is built consecutively with successive trees focusing on minimizing the loss compared to previous trees.

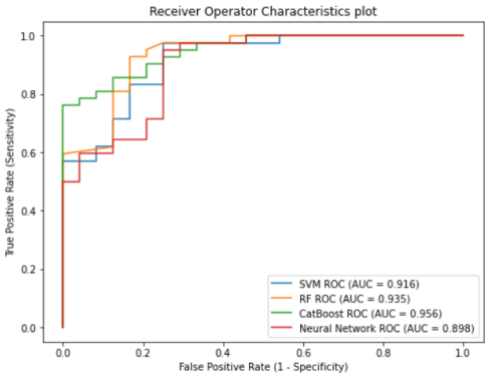


Figure 3: ROC plots for the 4 classification models considered in this work using the features produced by the RFE algorithm

Table 3
Comparing between the different classifiers

Model	Precision	Recall	F1-score	AUC
SVM	0.85	0.83	0.85	0.91
Random Forest	0.85	0.96	0.85	0.93
CatBoost	0.90	0.95	0.90	0.956
MLP	0.86	0.86	0.86	0.89

Compared with the help of SHAP algorithm, we can break down each prediction individually. As a demonstration, we took two individuals from the testing set: a 40-year-old patient that was predicted to have the disease and a 56-year-old patient that was not. In table 5, the non-normalized features of the two patients are shown.

Table 4
Non-normalized values for test patients 1 and 2

Feature Name	Patient 1 (positive for SLE)	Patient 2 (Negative for SLE)
AGE	40	56
Disease_Duration	21	13
PROV	Dakhiliyah	Muscat
Fever	N	N
ACL	N	N
Oral_Ul	N	N
Alopecia	N	N
Serositis	N	N
Renal	Nephritis	N

Proteinuria	Y	N
Neurologic	N	N
Hemolytic_Anemia	N	N
Leukopenia	N	N

The force plot attributes the positive prediction of patient 1 to renal disorders, and the patient's age (Figure 4.a). Since the values in Figure 4 are normalized we cross-reference them with table 4, we find that the patient is 40 which falls within the age group SLE is most active in. Additionally, the patient has been diagnosed with Lupus Nephritis a disease that is commonly caused by an auto-immune disorder. In contrast, patient 2 (Figure 4.b) displays a lack of any autoimmune manifestation, long disease duration, and the age of 56 makes him outside the age group that SLE is most active in.

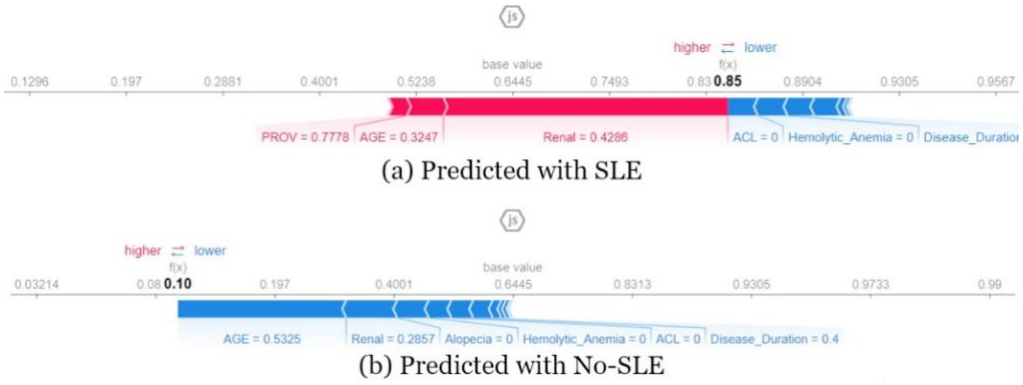
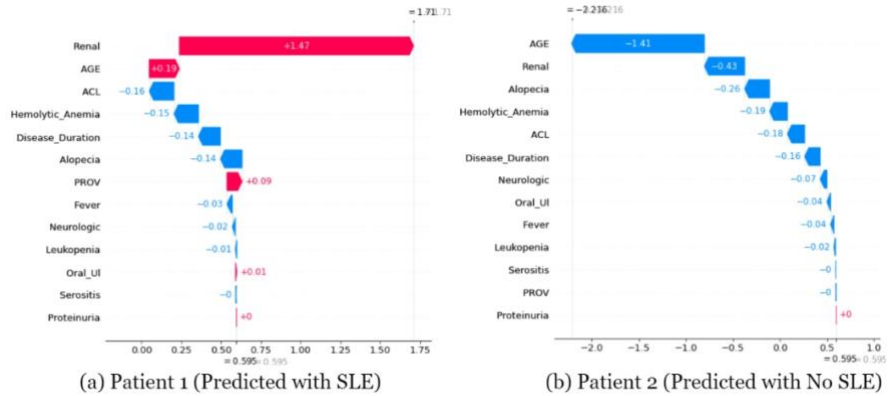


Figure 4: Force plot of CatBoost model prediction for patient 1 (values are normalized).

Looking at the waterfall plot in Figure (5.a), we find the feature with highest SHAP value for patient 1 is renal disorders 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, there were four blue features pushing the prediction probability lower toward class 0. The non-existence of alopecia, hemolytic anemia, and ACL in patient 1 profile in table 4 resulted in negative SHAP values. The remaining features had minimal impact on the prediction probability evident by their low SHAP values. The waterfall plot for patient 2 (Figure 5.b) indicates that age is the largest contribution toward class 0, followed by the absence of any renal disorders.

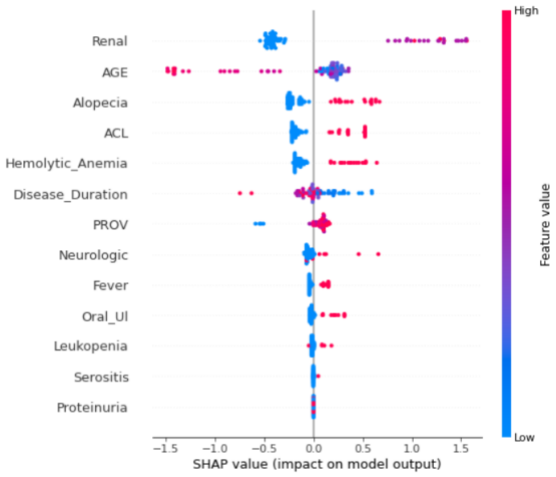


(a) Patient 1 (Predicted with SLE) (b) Patient 2 (Predicted with No SLE)

Figure 5: Waterfall plot of CatBoost model for patient 1. The waterfall plot displays SHAP values representing feature contribution toward a positive prediction.

Ranking Features. In Figure 6, 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. 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 late phases of the disease [15]. 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, Acute Cutaneous Lupus (ACL), and hemolytic anemia. The lowest contributing features are serositis, proteinuria, and leukopenia.



1.

Figure 6: Summery plot of CatBoost model. The summary plot combines feature importance with feature effects.

4. Conclusions

In this study, the first SLE prediction model has been developed with our proposed self-explainable framework that aims at establishing trust in ML prediction. SHAP interpretation tool was implemented to explain and justify individual predictions and thereby eliminate any risk of misclassification. Additionally, a minimum set of 13 early predictors achieved the highest scores of 0.95 AUC and 0.92 F1-score metrics. The dataset features comprise demographic and clinical symptoms available to physicians at early stages.

By interpreting Catboost predictions, we found that four clinical features had the highest influence on the prediction in addition to the patient's age. The features were alopecia, renal disorders, cutaneous lupus, and hemolytic anemia. All are considered 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 the Omani population.

With such scores, our model can predict with reasonable certainty the presence or absence of SLE. This can alert physicians to investigate further with the help of immunological tests such as antinuclear antibodies test and Anti-dsDNA test. Overall, our framework and its application can aid in providing a more practical introduction of machine learning and interpretation tools to medical diagnosis, thereby increasing the efficiency of medical testing and subsequently maximizing chances of disease mitigation and management. This is expected to reduce the cost, of medical care as well as decrease the cases of unmitigated severe cases of SLE.

References

[1] Nisengard R. Diagnosis of systemic lupus erythematosus. Importance of antinuclear antibody titers and peripheral staining patterns. Archives of Dermatology. 1975;111(10):1298-1300.

[2] Felten R, Lipsker D, Sibilia J, Chasset F, Arnaud L. The history of lupus throughout the ages. Journal of the American Academy of Dermatology. 2020;

[3] Piga M, Arnaud L. The Main Challenges in Systemic Lupus Erythematosus: Where Do We Stand?. Journal of Clinical Medicine. 2021;10(2):243.
[4] 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.
[5] 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. International Journal of Rheumatic Diseases. 2021;24(6):847-854.
[6] 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.
[7] Hancock J, Khoshgoftaar T. CatBoost for big data: an interdisciplinary review. Journal of Big Data. 2020;7(1).
[8] Anghel A, Papandreou N, Parnell T, et al. Benchmarking and Optimization of Gradient Boosting Decision Tree Algorithms. arXiv:180904559
[9] A. M, Brinks R, Dörner T, Daikh D, Mosca M, et al. European League Against Rheumatism (EULAR)/American College of Rheumatology (ACR) SLE classification criteria item performance. Annals of the Rheumatic Diseases. 2021;80(6):775-781.
[10] Eye A, Clogg C. Categorical variables in developmental research. San Diego: Academic Press; 1996.
[11] Prokhorenkova L, Gusev G, Vorobev A, et al. CatBoost: unbiased boosting with categorical features. arXiv:170609516
[12] Dorogush A, Ershov V, Gulin A. CatBoost: gradient boosting with categorical features support. arXiv:1810.11363
[13] Branco, P., Torgo, L., & Ribeiro, R. (2015). A survey of predictive modelling under imbalanced distributions. ArXiv:1505.01658
[14] Lundberg S, Lee S-I. A Unified Approach to Interpreting Model Predictions. arXiv:170507874
[15] 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. The Journal of Rheumatology. 2009;37(1):38-44.
[16] Binder A, Ellis S. When to order an antinuclear antibody test. BMJ. 2013;347(aug21 2):f5060-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 Pacific Journal of Allergy and Immunology. 2013;31(4).
[18] Arnaud L, Mathian A, Boddaert J, Amoura Z. Late-Onset Systemic Lupus Erythematosus. Drugs & Aging. 2012;29(3):181-189.
[19] Beckwith H, Lightstone L. Rituximab in Systemic Lupus Erythematosus and Lupus Nephritis. Nephron Clinical Practice. 2014;128(3-4):250-254.
[20] 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.
[21] Yu H, Nagafuchi Y, Fujio K. Clinical and Immunological Biomarkers for Systemic Lupus Erythematosus. Biomolecules. 2021;11(7):928.
[22] Giannouli S. Anaemia in systemic lupus erythematosus: from pathophysiology to clinical assessment. Annals of the Rheumatic Diseases. 2006;65(2):144-148.
[23] Nisengard R. Diagnosis of systemic lupus erythematosus. Importance of antinuclear antibody titers and peripheral staining patterns. Archives of Dermatology. 1975;111(10):1298-1300.