EAI Endorsed Transactions
on Pervasive Health and Technology
Research Article EAI.EU

Prediction of Paediatric Systemic Lupus Erythematosus Patients Using Machine Learning

Raja Rajeswari Ponnusamy[1,*], Lim Chun Cheak[2], Elaine Chan Wan Ling[3] and Lim Sern Chin[4]

[1]School of Computing, Asia Pacific University, 57000 Kuala Lumpur, Malaysia, raja.rajeswari@apu.edu.my
[2]School of Computing, Asia Pacific University, 57000 Kuala Lumpur, Malaysia, TP068620@mail.apu.edu.my
[3]International Medical University, 57000 Kuala Lumpur, Malaysia, elainechan@imu.edu.my
[4]Universiti Teknologi Mara, 40450 Shah Alam, Malaysia, sernchin@gmail.com

Abstract

Paediatric systemic lupus erythematosus (pSLE) is an autoimmune disease where the body's immune system attacks its own tissues, leading to organ damage. Advances in medical technology and the integration of artificial intelligence have significantly reduced the mortality rate of pSLE patients and improved their quality of life. Various studies have explored the link between environmental pollution and pSLE, utilizing machine learning to identify common gene expressions associated with the disease. However, the application of machine learning, particularly neural networks, to predict the status of pSLE patients over different timeframes remains underexplored. This study aims to demonstrate the effectiveness of support vector machines (SVMs) and neural networks in predicting the status of pSLE patients. Results show that without SMOTE balancing, both SVMs and neural networks achieved an accuracy of 68.09%, while neural networks achieved the highest accuracy of 77.78% after SMOTE balancing. Healthcare stakeholders can employ these machine learning techniques to provide early insights into patients' future health status based on their current condition, thereby improving patient outcomes.

Keywords: Neural Network, Support Vector Machine, PLSE.

Received on 11 02 2024, accepted on 28 05 2024, published on 20 06 2024

doi: 10.4108/eetpht.10.6386

1. Introduction

Systemic lupus erythematosus (SLE) is a common type of lupus and autoimmune disease. In individuals diagnosed with SLE, the immune system mistakenly attacks their own tissues due to the large deposit of antinuclear auto-antibodies (ANA). These ANA confuse normal proteins with foreign or harmful entities, triggering a cascade of inflammation and self-attack [1]. This autoimmune response can lead to widespread inflammation and potential damage to organs such as the skin, brain, joints, lungs, kidneys, and blood vessels. Additionally, SLE patients face a higher risk of cancer, infections, bone tissue death and pregnancy complications [2].

SLE patients can experience functional impairment in physical, mental, and social aspects of their lives, significantly impacting their quality of life, especially when they suffer from fatigue, one of the most common symptoms of SLE. Although there is currently no remedy for lupus, medical interventions and lifestyle changes can aid in managing SLE by minimizing disease activity, ensuring long term survival, avoiding organ damage, reducing drug toxicity and improving patients' quality of life. Conversely, complications may worsen, and risk of death may increase if patients lack access to care,

*Corresponding author. Email: raja.rajeswari@apu.edu.my

ineffective treatments and diagnosis, or demonstrate poor compliance with therapeutic regimens [2].

The exact factors causing SLE are not clearly understood but may be linked to genetic, environmental, hormonal, and certain medication-related factors. Age, gender, and race are among the significant risk factors. The epidemiology of pediatric systemic lupus erythematosus (pSLE) indicates that the median age of onset is between 11 to 12 years, with cases being quite rare under the age of 5. Statistics show an incidence of 0.3-0.9 per 100,000 children-years and a prevalence of 3.3-8.8 per 100,000 children. Women appear to be more susceptible: approximately 80% of pSLE patients are female, and the condition is more frequently reported amongst Asians, African Americans, Hispanics and native Americans [3]. Additionally, it was noted that geographic variables could influence SLE morbidity, even among patients with the same sex and ethnicity [4]. Studies indicate different complications in SLE patients depending on their region, such as coastal, mountainous, or jungle areas, due to different exogenous factors. [1]

Over the previous 30 years, more than 10,000 SLE cases have been diagnosed in Malaysia, though this figure may underestimate the true prevalence. According to the Malaysian SLE Association, there are likely many more undiagnosed cases within Malaysia [5]. Given that childhood-onset SLE (pSLE) may be more severe than adult-onset SLE (aSLE) due to differences in clinical and biological features [6], this research aims to develop a predictive model using a deep learning neural network to identify pSLE patients early, providing them with comprehensive treatments to alleviate their suffering in Malaysia.

Malaysia is a racially heterogeneous country with Malays, Chinese and Indians comprising the largest three majority groups, alongside several minority Bumiputera races such as Iban, Kadazan and Bayau. Research indicates that Chinese communities are more susceptible to SLE in comparison with Malays and Indians [7]. Alhough most of the genes are not responsible for pSLE susceptibility in Malaysia, distinct ancestral genotypes affecting the development or progression of pSLE may contribute to genetic variability among different racial groups [8].

Several retrospective cohort analyses in Malaysia have explores pSLE's unknown disease features and outcomes. According to Lim, the female-to-male ratio of patients is 7.3:1, with age specific ratios of 2.3:1, 8.3:1 and 15:1 for ages 0-5, 6-12 years, and 13-18 years respectively [9]. Fever, vasculitic rash and fatigue are among the most frequent clinical features, with renal and central nervous system (CNS) damages being major organ involvements. Renal disorder is most common among pSLE patients, followed by malar rash, oral ulcers, prolonged fever and arthritis [10].

One study found that pSLE patients had higher prevalence and mortality rates than aSLE patients due to acute pancreatitis [11]. Systemic Lupus Erythematosus Disease Activity Index (SLEDAI) scores were higher in pSLE patients with acute pancreatitis compared with aSLE patients (21.77 versus 13.37), and multivariate logistic regression analysis proved that acute pancreatitis was the most significant risk factor for mortality. Besides, Furthermore, neuropsychiatric lupus (NPSLE) has been given as a marker of severe disease in pSLE patients, who present a myriad of clinical features, such as seizures, delirium, and visual complaints [12].

With the emergence of machine learning and deep learning in medical applications, these techniques were widely used in aiding the diagnosis of pSLE. Researchers implemented these techniques to predict the presence of certain symptoms among pSLE patients, have attempted to characterize the immune cell profile of pSLE patients and investigate the relationship of disease trajectory with time [13][14]. However, predictive methodologies are yet to be used in predicting disease outcome and the long-term status of patients using medical features and records over short periods. The study proposes using artificial neural networks to generate state- of-the-art predictions for multiclass targets in pSLE patients.

2. Methods

This study uses a secondary dataset obtained from the International Medical University (IMU), which was collected from existing pSLE patients. The dataset includes 141 medical records with 14 features describing patients' medical history and symptoms. These features encompass demographic factor of patients such as age, gender and race, clinical features upon diagnosis, Systemic Lupus International Collaborating Clinics (SLICC) classification criteria for SLE, disease progression with SLEDAI scores over a 10-years period, disease damage over a 10-year period, renal disease treatment, non-renal flares and status of patients. The status is the target variable, while the other features serve as predictors. The status class is multiclass, including remission on medications, remission without medications, minimal disease activity, chronic active disease, and death. As classes in status are imbalanced, the SMOTE resampling technique is used to generate synthetic samples for the underrepresented classes. This approach enhances the performance of the neural network model by providing more instances for the training set.

2.1. Support Vector Machine (SVM)

SVM is a supervised learning algorithm that performs classification tasks by identifying a hyperplane that best separates the two classes in n-dimensional space. For linear separation, this hyperplane usually acts as a linear decision boundary that maximizes the distance between the two classes of data points. Although SVM only supports binary classification while the dataset used contains 5 different classes, the algorithm can perform multiclass classification by breaking down the

multiclassification problem into multiple binary classification problems using the One-vs-Rest approach [15].

SVM was selected as the benchmark model for this project due to its strong performance in classification tasks and its ability to model non-linear decision boundaries by changing the kernel parameter. Various kernels are available to enhance performance, including radial basis function (RBF), polynomial function and Gaussian function [16]. Thus, SVM is considered a suitable model to compare against the proposed neural network model.

2.2. Neural Network

Neural networks, a subset of machine learning and the core of deep learning, are designed to mimic the signal transmission between neurons in the human brain. The ability to handle unstructured data, such as texts and images, has enabled their widespread application in daily life, including chatbots, recommendation engines and image classification [17]. Typically, a neural network consists of one input layer, multiple hidden layers, and one output layer. The input layer receives the data, with the number of input nodes determined by the number of features. The hidden layers perform computations and extract features from the data, with the number of layers and nodes varying based on the dataset's complexity. The output layer processes the input from hidden layers to make a final prediction. After model implementation, the predictions on the training and test sets are evaluated by accuracy to assess their fit and check for overfitting or underfitting issues. Accuracy serves as a fundamental metric for performance comparison against benchmarks and for evaluating each experiment's outcomes.

3. Result and Discussions

In this section, the parameters and results of each experiment are further described and tabulated (Table 1). The architectures of the neural network for each experiment will also be displayed.

Experiment I - Support Vector Machine
Two different kernels were selected for the SVM experiments: the RBF (radial basis function) and polynomial functions. For SVM with the RBF kernel, the best results were obtained with a C value of 0.1 and a gamma value of 0.5. The hyperparameter C defines the error control of SVM, where a low C value indicates low error and vice versa. Gamma, specific to the RBF kernel, determines the curvature of the decision boundary, with a higher gamma resulting in more curvature. For SVM with the polynomial kernel, the default degree value of 3 was used, along with a C value of 0.1.

Experiment II - Neural Network
In the first neural network model trained with all features, the best result was obtained with one dense input layer, two dense hidden layers and one dense output layer. The input layer and two hidden layers each contained 25 nodes, utilizing the “Relu” activation function, while the output layer was comprised of 5 nodes corresponding to the number of classes in the dataset, and employed the “SoftMax” activation function. To mitigate overfitting, dropout layers with a rate of 0.25 were alternately added between the input layer and the two hidden layers, which helped drop a fraction of input units. Given the multiclass nature of the target, categorical cross entropy was used as the loss function. For optimizers, adaptive moment estimation (adam) with a 0.001 learning rate was used, and the number of epochs was set to 20.

Experiment III & IV- Neural Network
The second and third neural network model experiments used a modified dataset which removed features related to disease activity and progression beyond the first to fifth years. Optimal accuracy was obtained for both experiments when the number of nodes in the input and hidden layers was reduced to 20, with dropout layers (rate = 0.25) inserted between the two hidden layers. The activation function, loss function, learning rate and type of optimizer remained consistent with previous experiments.

Experiment V - Neural Network (Balanced dataset)
To address class imbalance, a balanced dataset was created by using SMOTE sampling, which increased the number of instances in the underrepresented classes. Specifically, the "remission without medication" and "death" classes were augmented from 2 and 1 instances, respectively, to 20 each, while other classes remained unchanged. Post-balancing and data splitting, the data was passed to a neural network model consisting of one dense input layer, two dense hidden layers and one output layer. Optimal accuracy was achieved with 50 nodes in the input and hidden layers, without incorporating dropout layers.

Table 1. Accuracy result to predict pSLE status.

Experiment	Accuracy (%)
Decision Tree Classifier [18]	58.14
Random Forest Classifier [18]	58.14
RBF SVM	68.09
Gaussian SVM	61.7
Neural network (All features)	68.09
Neural network (disease severity with 1st year onwards removed)	61.7
Neural network (disease severity with 5th year onwards removed)	61.7

Neural network (balanced dataset)	**77.78**

From Table 1, it is evident that all experiments outperformed the best accuracy of previous work, likely due to different data preprocessing techniques. Notably, the polynomial SVM in Experiment I performed on par with the neural network model in Experiment II, while the RBF SVM matched the accuracy of the neural network model in Experiment III. Despite eliminating different ranges of features in Experiments III and IV (one dataset excluded disease progression and severity from the first year onwards, while the other excluded from the fifth year onwards), both datasets yielded the same accuracy in neural network model. In Experiments III and IV was lower compared to Experiment II.

Experiment V achieved the highest accuracy generated with a neural network model trained on a balanced dataset that has more instances of "remission without medication" and "death" classes. This indicates the neural network's effectiveness in pSLE multiclass classification and its potential for early prediction for patients' future status. Early prediction of pSLE status could serve as reference for healthcare workers to anticipate severe complications and provide better medical attention to guide patients towards a healthier recovery.

The neural network performed best on the balanced dataset with more observations in rare classes, such as "remission without medication" and "death", demonstrating its ability to predict well when it can study ample samples from each class during model training. However, the SMOTE technique's synthetic samples may not fully reflect patients' conditions. Moreover, the neural network's performance stagnanted or decreased with imbalanced datasets, despite parameter adjustments aimed at increasing accuracy.

Hence it is suggested that future research should focus on experiments with balanced datasets containing real patient samples rather than synthetic ones. While patient data is limited due to the rare occurrence of pSLE and improved medical technology improved reducing early stage mortality, data could be aggregated from multiple hospitals or countries to create a more comprehensive dataset. Additionally, considering features such as the pollution level of patients' living environments could could provide insights into the impact of these factors on disease severity and patient status.

4. Conclusions

In summary, SVMs and neural networks have proven effective tools to predict the status of pSLE patients from features such as demographic factors and disease progression and severity. Without class balancing, SVM has performed on a par with the neural network benchmark, suggesting its suitability depending on the scale of a dataset. As datasets expand, neural networks will become a better option. The highest accuracy was obtained when the neural network was trained on a balanced dataset with increased instances of synthetic rare classes, underscoring the importance of retrieving real records of patients from multiple sources to enhance neural network prediction. Features such as pollution levels in patients' living environments should be considered, as these have been proven to affect pSLE patient health according to previous studies.

Acknowledgements.

This research funded by Asia Pacific University of Technology and Innovation, Malaysia with research grant of RDIG/06/2020.

References

[1] Angel Chamorro Quijano, S., Muñoz Melgarejo, M., Rodríguez, G., Marlene Muñoz Saenz, D., Caroline Muñoz Saenz, J. Analysis of the relationship of systemic lupus erythematosus with exogenous factors in Peru. 2021. *4th International Conference on Digital Medicine and Image Processing*. 72-76.

[2] CDC. *Systemic lupus erythematosus (SLE)*. Centers for Disease Control and Prevention. 2022. https://www.cdc.gov/lupus/facts/detailed.html#:~:text=doing%20about%20SLE%3F,What%20is%20SLE%3F,%2C%20kidneys%2C%20and%20blood%20vessels.

[3] Levy, D. M., Kamphuis, S. Systemic lupus erythematosus in children and adolescents. *Pediatric Clinics of North America*. 2021 *59*(2), 345–364. https://doi.org/10.1016/j.pcl.2012.03.007

[4] Singh, R. R.,Yen, E. Y. Sle mortality remains disproportionately high, despite improvements over the last decade. *Lupus*, 2018. *27*(10), 1577–1581.

[5] PSLEM. *What is Sle?* 2022. https://lupusmalaysia.org/en/what-is-sle

[6] Descloux, E., Durieu, I., Cochat, P., Vital-Durand, D., Ninet, J., Fabien, N., Cimaz,R. Influence of age at disease onset in the outcome of paediatric systemic lupus erythematosus. *Rheumatology*. 2009. *48*(7), 779–784. https://doi.org/10.1093/rheumatology/kep067

[7] Chai, H. C., Phipps, M. E., Chua, K. H. Genetic risk factors of systemic lupus erythematosus in the Malaysian population: A Minireview. *Clinical and Developmental Immunology.* 2012, 1–9. https://doi.org/10.1155/2012/963730

[8] Lee, H.-S., Bae, S.-C. What can we learn from genetic studies of systemic lupus erythematosus? implications of genetic heterogeneity among populations in SLE. *Lupus*. 2010. *19*(12), 1452–1459. https://doi.org/10.1177/0961203310370350

[9] Lim, S. C., Chan, E. W.,Tang, S. P. Clinical features, disease activity and outcomes of Malaysian children with paediatric systemic lupus erythematosus: A cohort from a tertiary centre. *Lupus*. 2020. 29(9), 1106–1114.

[10] Nazri, S. K., Wong, K. K., Hamid, W. Z. Pediatric systemic lupus erythematosus. *Saudi Medical Journal*. 2018. *39*(6), 627–631. https://doi.org/10.15537/smj.2018.6.22112

[11] Blaskievicz, P. H., Silva, A. M., Fernandes, V., Junior, O. B., Shimoya-Bittencourt, W., Ferreira, S. M., da Silva, C. A. Atmospheric pollution exposure increases disease activity of systemic lupus erythematosus. *International*

Journal of Environmental Research and Public Health. 2020. *17*(6), 1984.

[12] Lim, S. C., Yusof, Y. L., Johari, B., Kadir, R. F., Tang, S. P. Neuropsychiatric lupus in Malaysian children: Clinical characteristics, imaging features and 12-month outcomes. *The Turkish Journal of Pediatrics*. 2021.*63*(5), 743.

[13] Rajimehr, R., Farsiu, S., Kouhsari, L. M., Bidari, A., Lucas, C., Yousefian, S., Bahrami, F. Prediction of lupus nephritis in patients with systemic lupus erythematosus using artificial neural networks. *Lupus*. 2002. *11*(8), 485–492.

[14] Robinson, G. A., Peng, J., Dönnes, P., Coelewij, L., Naja, M., Radziszewska, A., Wincup, C., Peckham, H., Isenberg, D. A., Ioannou, Y., Pineda-Torra, I., Ciurtin, C., Jury, E. C. Disease-associated and patient-specific immune cell signatures in juvenile-onset systemic lupus erythematosus: Patient stratification using a machine-learning approach. *The Lancet Rheumatology. 2020. 2*(8).

[15] Baeldung. *Multiclass classification using support Vector Machines*. Baeldung on Computer Science. 2022. https://www.baeldung.com/cs/svm-multiclass-classification

[16] goelaparna1520. *Support Vector Machine in machine learning*. GeeksforGeeks. 2023. https://www.geeksforgeeks.org/support-vector-machine-in-machine-learning/

[17] Nicholson, C. *A beginner's Guide to Neural Networks and deep learning*. Pathmind. 2022. https://wiki.pathmind.com/neural- network

[18] Singh, J. K. J., Ponnusamy, R. R., Ling, E. C. W., Chin, L. S. Early Prediction of Lupus Disease: A Study on the Variations of Decision Tree Models. *Advances in Bioengineering and Biomedical Science Research*. 2022. 5(4).