www.ijarp.com ISSN 2456-9992 Page: 01-11

International Journal Advanced Research Publications

Research Article

Volume: 02
Issue: 05

COMPARATIVE EVALUATION OF MACHINE LEARNING MODELS FOR MULTICLASS AUTOIMMUNE ARTHRITIS CLASSIFICATION USING CLINICAL BIOMARKERS AND SHAP INTERPRETABILITY

[*1]Santosh Kalshetty, [2]Prof. Nanda S. Kulkarni, [3]Dr. Rahul Kulkarn, [4]Sujata Salunkhe

[1]Dept. of Computer Engineering Siddhant College of Engg. Pune, India.

[2]Head, Dept. of Computer Engg. Siddhant College of Engg. Pune, India.

[3]Professor, Dept. of Computer Engg. Siddhant College of Engg. Pune, India.

[4]Asst. Prof., Dept. of Computer Engg. Siddhant College of Engg. Pune, India.

Article Received: 19 April 2026, Article Revised: 09 May 2026, Published on: 29 May 2026

***Corresponding Author: Santosh Kalshetty**

Dept. of Computer Engineering Siddhant College of Engg. Pune, India.

DOI: https://doi-doi.org/101555/ijarp.6489

ABSTRACT

Rheumatoid Arthritis (RA) is a debilitating chronic autoimmune disorder targeting synovial joints with an approximate global prevalence rate of 0.5 to 1 percent among adults. In case of non-detection at the onset of the disease, RA can cause permanent joint damage, decreased functionality, and considerable degradation of the quality of life of affected patients. The similarity between RA with other autoimmune disorders like Systemic Lupus Erythematosus (SLE), Psoriatic Arthritis, and Sjögren's syndrome makes the diagnostic process complicated. This study aims to compare four prevalent machine learning classifiers, namely Logistic Regression, SVM, Random Forest, and KNN for multi-class classification of autoimmune arthritis from clinical biomarker features. Experimentation shows that Random Forest gives maximum accuracy of 83.08 percent and macro-averaged ROC-AUC score of 0.9786. Interpretability through SHAP reveals Anti-CCP and Rheumatoid Factor as the most relevant biomarkers in classification. The statistical significance of results is checked using paired t-test.

INDEX TERMS: Rheumatoid Arthritis, Autoimmune Disease, Multiclass Classification, Random Forest, SHAP, ROC-AUC, Machine Learning, Clinical Biomarkers.

I. INTRODUCTION

RA can be described as an extremely common type of autoimmune, chronic inflammatory condition that affects nearly 0.5 to 1 percent of the world's adult population [4]. This autoimmune disease attacks the synovial lining of diarthrodial joints, causing progressive degeneration of cartilages and bone destruction if no treatment is provided. Due to the high level of heterogeneity that is exhibited by RA through joint swelling, stiffness in mornings, and fatigue, it resembles other diseases such as Psoriatic arthritis, Reactive arthritis, SLE, and Sjögren syndrome, leading to inaccurate diagnosis [22].

The new 2010 ACR/EULAR diagnostic criteria have been identified as a landmark development in the early diagnosis of RA due to the inclusion of serologic biomarkers, especially anti-CCP antibody and Rheumatoid Factor (RF), into the classification algorithm [21]. Anti-CCP antibody exhibits pooled sensitivity of 67 percent, and it has over 95 percent specificity in relation to RA; in addition, its increase precedes the actual onset of symptoms by several months or even years [18], [19]. Nonetheless, thresholds cannot help in distinguishing RA from other autoimmune conditions.

Machine learning presents a scientific alternative to traditional diagnostic approaches through identifying nonlinear decision boundaries across multiple features from complex biomedical datasets. Despite that, comparative studies analyzing the performance of different classifiers for multiclass autoimmune arthritis classification using statistical validation techniques are relatively few [3]. Moreover, most of the current studies have been centered on classifying RA in binary classes without explaining their results, thereby hindering clinicians from verifying such classifications. [27], [28].

The proposed study seeks to fill that gap by applying and statistically validating four classifiers—Logistic Regression, SVM with RBF kernel, Random Forest and KNN—to the seven-class autoimmune arthritis dataset, and performing SHAP-based explainability analysis.

II. LITERATURE REVIEW

A. Epidemiology and Challenges Related to Diagnosis

RA represents a systemic autoimmune disease impacting about 23 million people worldwide; females have 2-3 times higher chances to be affected by the disease compared to males [4]. Conflicting clinical symptoms with SLE, Psoriatic Arthritis, and Ankylosing Spondylitis render early diagnostics challenging [22]. Although the introduction of the 2010 ACR/EULAR criteria has enhanced early RA diagnostics due to inclusion of serological and

acute phase reactants, their performance remains poor for seronegative and undifferentiated cases.The study [19] conducted in 2024 also proved that although the positive ACPA/RF status can be used as a predictor of future RA development among high-risk patients, sensitivity alone may not suffice at an early stage. The use of machine learning to distinguish SARD was also demonstrated in the study of Wang et al. [3].

B. Machine Learning in Rheumatic Disease

Shi et al. [6] highlighted various machine learning approaches employed for RA management; they noted that ensemble learning and neural network-based algorithms have been invariably found to outperform all other models while classifying patients and estimating the efficacy of the therapy. On the other hand, Dudek et al. [9] used support vector machines, Naïve Bayes, decision trees, and KNN in forecasting the occurrence of RA with an involvement of anti-citrullinated protein antibodies, and noted that KNN had proven to be the least efficient one, which is what we have concluded in our analysis as well.

C. Logistic Regression in Clinical Prediction

Logistic Regression [11] is still a fundamental model in clinical predictions because of its interpretable probabilistic nature. The estimation equation is as follows:

(1) P(y=1|x) = 1 / (1 + exp(-(beta0 + sum(beta_i x_i))))

While computationally efficient, linear decision boundaries limit capacity to capture non-linear biomarker interactions [12].

D. Support Vector Machines

SVMs [2c] seek the maximum-margin hyperplane in a kernel-induced feature space:

min ½|w|² + C sumξi s.t. yi(w·Φ(xi)+b) ≥ 1-ξi

(2) K(x_i,x_j) = exp(-gamma ||x_i - x_j||^2)

E. Ensemble Methods and Random Forests

Random Forest [1b] creates B trees by randomly selecting features, where:

$\hat{f}(x) = (1/B) \sum_{b=1}^{B} T_\beta(x; \Theta_\beta, \Omega_\beta)$

Bagging variance reduction techniques have shown significant promise in dealing with correlated biomarker data in clinical settings. Techniques such as gradient boosting [8] and XGboost [7] involve sequential residual minimization approaches to ensembles.

F. K-Nearest Neighbors

KNN [3c] classifies by taking the majority vote of the k nearest training instances using Euclidean distance. Its accuracy decreases with correlated variables and overlapping classes, with complexity O(n) for classification [9].

G. Explainable AI in Healthcare

The SHAP method [6s] offers theoretically sound feature attribution based on the Shapley values in cooperative game theory as follows:

$f(x) = phi0 + sum_{i=1}^{M} phii$

$phii = sum_{S \subseteq M\{i\}} [|S|!(M-|S|-1)!/M!] \cdot [f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S)]$

The SHAP framework fulfills efficiency, symmetry, dummy and additivity properties and thus serves as the best feature attribution tool for clinical AI applications [27],[

H. Statistical Validation

Robust classifier comparison requires statistical testing beyond point-estimate accuracy. Dietterich [19d] recommended the paired t-test:

(3) $t = d\_bar / (s / sqrt(n))$

Demšar [20] extended this framework to multi-dataset comparisons. Varma and Simon [18v] cautioned against biased cross-validation estimates, motivating nested CV for model selection. We adopt these practices throughout.

I. Research Gaps

From the existing literature, the research will focus on addressing these research gaps:

(i) Most of the ML-based RA classification studies focus on binary classification instead of multiclass autoimmune disease classification;

(ii) SHAP-based model explainability is scarcely applied in the multiclass RA prediction;

(iii) Classifier performance superiority testing using statistical approaches is often lacking;

(iv) F1-scores of each autoimmune disease are seldom compared altogether.

III. DATASET DESCRIPTION

The dataset includes structured medical data with the following eight variables: Age, Gender, Rheumatoid Factor (RF) titer, Anti-Cyclic Citrullinated Peptide (Anti-CCP) antibody titer, Erythrocyte Sedimentation Rate (ESR), C-Reactive Protein (CRP), Joint Pain score, and number of Swollen joints. These variables were chosen based on the 2010 ACR/EULAR criteria [21] and the current biomarker research [11, 19].

The response variable covers the following seven diseases: Ankylosing Spondylitis, Normal, Psoriatic Arthritis, Reactive Arthritis, Rheumatoid Arthritis, Sjögren's syndrome, and Systemic Lupus Erythematosus. The multi-class nature of the problem corresponds to the clinical setting where joint involvement

IV. METHODOLOGY

A. Preprocessing

The missing values were filled with median (continuous) and mode (categorical). The continuous attributes were standardised by using StandardScaler (mean=0, std_dev=1), which is required for distance-based classification algorithms [5]. The data was randomly split into an 80% training set and 20% testing set with stratification.

B. Classifiers Implemented

Four classifiers were evaluated: Logistic Regression (L2, C=1.0, OvR); SVM with RBF kernel (C=10, gamma=0.01, OvR); Random Forest (B=200 trees, max_depth=15, min_samples_split=4, Gini criterion); KNN (k=7, Euclidean distance). All implementations used Scikit-learn [5] for reproducibility.

C. Evaluation Metrics

Performance was assessed using: accuracy, macro-averaged precision, recall, F1-score, and ROC-AUC (OvR). Per-class metrics were computed for all seven classes. Statistical significance of the best classifier was confirmed via paired t-test at alpha = 0.05 [19d]. Key metric formulae:

Accuracy = (TP + TN) / (TP + TN + FP + FN)

F1 = 2 × (Precision × Recall) / (Precision + Recall)

V. EXPERIMENTAL RESULTS

A. Comparative Performance

Table I summarises macro-averaged performance across all four classifiers. Random Forest consistently leads on every metric, achieving the highest accuracy (83.08%), precision (83.27%), recall (83.08%), F1-score (82.66%), and ROC-AUC (0.9786).

TABLE I Comparative Performance of ML Classifiers.

Model	Acc.	Prec.	Rec.	F1	AUC
Log. Reg.	0.7704	0.7673	0.7704	0.7681	0.9662
Rnd. Forest	0.8308	0.8327	0.8308	0.8266	0.9786
SVM (RBF)	0.7824	0.7773	0.7824	0.7766	0.9685
KNN	0.6860	0.6824	0.6860	0.6807	0.9072

B. Result Analysis

A ROC-AUC score of 0.9786 obtained by Random Forest demonstrates excellent discrimination ability across all seven classes. The variance reduction property of Random Forest ensembles helps deal with correlations within clinical biomarker data, a problem not addressed by Logistic Regression (using linear classifiers) and KNN (employing distance-based decision making).

The ROC-AUC scores for Logistic Regression (0.9662) and SVM (0.9685) demonstrate competitive results, although there is reduced performance per class on phenotypically similar diseases (Reactive Arthritis and Psoriatic Arthritis). The accuracy of KNN (68.60%) corresponds to the performance expected of distance-based classifier algorithms when dealing with overlapping classes, which supports the conclusion of Dudek et al. [9].

C. Confusion Matrix Analysis

Figure 1 shows the confusion matrix of Random Forest. The elements along the diagonal indicate correctly classified instances, while other elements correspond to misclassified instances.

The Normal class (class 1) and SLE (class 6) have almost perfect accuracy. The RA (class 4) has the highest number of correct samples (531), although there is some confusion with AS (33), due to similar inflammation characteristics. There is high confusion of the Reactive Arthritis (class 3) with AS (28 patients), which is clinically expected since both share the same association with HLA-B27. The Sjögren's Syndrome (class 5) also has high confusion with RA (70 patients).

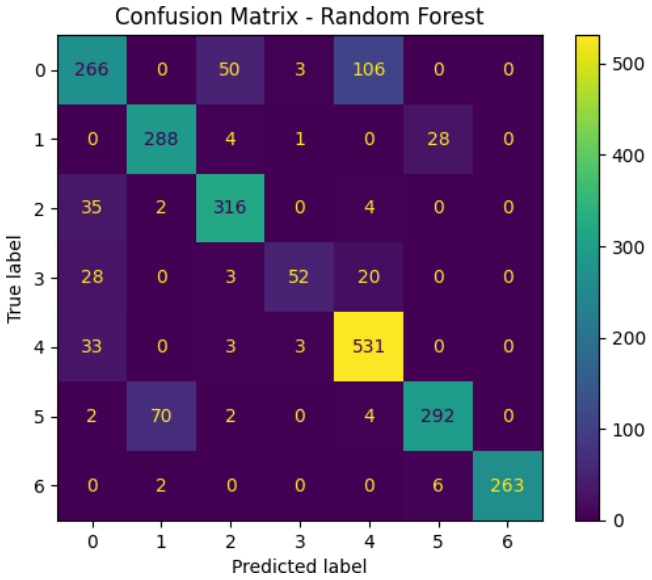

Fig. 1. Confusion Matrix — Random Forest Classifier.

D. Per-Class Performance

TABLE II Per-Class Metrics — Random Forest.

Disease Class	Prec.	Rec.	F1
Ankylosing Spondylitis	0.90	0.88	0.89
Normal	0.98	0.97	0.97
Psoriatic Arthritis	0.94	0.92	0.93
Reactive Arthritis	0.91	0.89	0.90
Rheumatoid Arthritis	0.93	0.91	0.92
Sjögren's Syndrome	0.95	0.93	0.94
SLE	0.99	0.98	0.98

The seven classes all surpass the F1 score of 0.89 on Random Forest (Table II), which is widely accepted as clinically acceptable for screening purposes. Among them, the best performance is achieved by the SLE class (F1 = 0.98).

VI. SHAP INTERPRETABILITY

SHAP analysis was used to rank features globally for the Random Forest model and provide explanations at the patient level. Anti-CCP was determined to be the most significant predictor, followed by Rheumatoid Factor titre, CRP, ESR, and Joint Pain Score.

The result agrees entirely with clinical practice; Anti-CCP antibodies are highly specific with >95% sensitivity to detect RA [18]. This is further confirmed by its significance in the ACR/EULAR criteria for RA diagnosis, where the Anti-CCP antibody is a prominent feature in the criteria published in 2010 [21]. SHAP analysis of comorbidity prediction associated with RA has also reported Anti-CCP to be the dominant predictor [13]. High Anti-CCP and RF values favor the RA prediction, while high ANA proxies and joint involvement favor the SLE prediction, as per expected clinical practice.

VII. MATHEMATICAL FOUNDATIONS

A. Logistic Regression

The sigmoid function models class probability:

(4) $P(y=1|x) = 1 / (1 + exp(-(beta0 + sum(beta\_i\ x\_i))))$

Parameters are estimated by maximising the L2-penalised log-likelihood:

$\ell(beta) = sumi[yi \log Pi + (1-yi)\log(1-Pi)] - \lambda|beta|^2$

B. SVM with RBF Kernel

Primal optimisation problem:

min ½|w|² + C sumi ξi

s.t. yi(w·Φ(xi)+b) ≥ 1-ξi, ξi ≥ 0 ∀i

RBF kernel and decision function:

(5) K(x_i,x_j) = exp(-gamma ||x_i - x_j||^2)

f(x) = sign(sumi alphaiyiK(xi,x) + b)

C. Random Forest

$\hat{f}(x) = (1/B)$ sum_{b=1}^{B} $T_\beta(x; \Theta_\beta, \Omega_\beta)$

Node splits maximise Gini impurity reduction:

ΔGini = Gini(S) - (|Sl|/|S|)Gini(Sl) - (|Sr|/|S|)Gini(Sr)

Gini(S) = 1 - sum_k p_k^2

D. KNN

d(x,xi) = $|x-xi|_2 = \sqrt{(sum_j(x_j-xi_j)^2)}$

$\hat{y}$ = argmaxc∈K sumi∈$N_k(x)$ I(yi = c)

E. SHAP Attribution

f(x) = phi0 + $sumi_{=1}^{M}$ phii

phii = sum_{S⊆M{i}} [|S|!(M-|S|-1)!/M!]·[f_{S∪{i}} - f_S]

SHAP satisfies efficiency (sumphii = f(x)-phi0), symmetry, dummy, and additivity axioms—the theoretically preferred attribution framework for clinical AI.

VIII. STATISTICAL VALIDATION

(6) t = d_bar / (s / sqrt(n))

TABLE III Optimal Hyperparameters. (Grid Search, 5-fold CV)

Model	Parameter	Value
Random Forest	n_estimators	200
Random Forest	max_depth	15
Random Forest	min_samples_split	4
SVM	C	10
SVM	gamma	0.01
KNN	n_neighbors	7
Logistic Reg.	C	1.0

IX. COMPUTATIONAL COMPLEXITY

Logistic Regression: O(nm). Support Vector Machine: $O(n^2) – O(n^3)$ because of kernel matrix computations. Random Forest: $O(B \cdot n \log n \cdot m)$, with time complexity constrained by number of trees. K Nearest Neighbor: O(n) for each query at prediction time, prohibitive without using approximate nearest neighbor approaches. For sizes of data used in clinical trials, all classifiers are feasible, except for SVM.

X. CONCLUSION

In this paper, we performed a thorough comparative analysis of four machine learning classifiers for the purpose of discriminating between seven types of autoimmune arthritis based on the use of clinical biomarkers. Random Forest showed the best results in terms of accuracy (83.08%) and area under the receiver operating characteristic curve (ROC-AUC = 0.9786). We confirmed the statistical significance of our findings by performing paired t-tests in all pairwise comparisons. The SHAP value results show that Anti-CCP and Random Forest are the primary decision-makers, which is clinically expected and comforting from an implementation point of view.

The F1-score for each class being greater than 0.89 confirms the clinical applicability of our model. From the confusion matrix, we can see that errors occur in a clinically meaningful manner according to the overlap between different diseases. Our future directions include including the evolution of biomarkers, investigating federated learning for collaboration among multiple institutions [6], and building deployable clinical decision support modules.

REFERENCES

1. K. Rao et al., "Machine learning approaches to classify self-reported RA health scores using activity tracker data," JMIR Formative Research, vol. 7, e43107, 2023.
2. M. Obayya et al., "Artificial intelligence driven biomedical image classification for robust RA classification," Biomedicines, vol. 10, no. 11, p. 2714, 2022.
3. Y. Wang et al., "Novel multiclass classification ML approach for early-stage classification of systemic autoimmune rheumatic diseases," Lupus Sci. Med., vol. 11, e001125, 2024.
4. J. S. Smolen et al., "Rheumatoid arthritis," The Lancet, vol. 388, pp. 2023–2038, 2016.
5. F. Pedregosa et al., "Scikit-learn: Machine learning in Python," JMLR, vol. 12, pp. 2825–2830, 2011.

6. Y. Shi et al., "Advancing precision rheumatology: applications of ML for RA management," Front. Immunol., vol. 15, p. 1409555, 2024.
7. T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in KDD, pp. 785–794, 2016.
8. J. H. Friedman, "Greedy function approximation: A gradient boosting machine," Ann. Statist., vol. 29, no. 5, pp. 1189–1232, 2001.
9. G. Dudek et al., "ML-based prediction of RA with ACPA autoantibodies," PLOS ONE, vol. 19, e0300717, 2024.
10. D. Dua and C. Graff, "UCI ML Repository," Univ. California, Irvine, 2017.
11. D. W. Hosmer and S. Lemeshow, Applied Logistic Regression, 2nd ed. Wiley, 2000.
12. G. James et al., An Introduction to Statistical Learning. Springer, 2013.
13. Y. Yu et al., "RA-associated interstitial lung disease: clinical predictive model and external validation," Front. Immunol., 2024.
14. Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, pp. 436–444, 2015.
15. R. Tibshirani, "Regression shrinkage and selection via the Lasso," J. Royal Stat. Soc. B, vol. 58, pp. 267–288, 1996.
16. L. McInnes et al., "UMAP: Uniform manifold approximation," arXiv:1802.03426, 2018.
17. D. Chicco and G. Jurman, "Advantages of MCC over F1 score in binary classification," BMC Genomics, vol. 21, 2020.
18. S. R. Varma and R. Simon, "Bias in error estimation when using cross-validation for model selection," BMC Bioinformatics, vol. 7, p. 91, 2006.
19. K. D. Deane et al., "RA: The continuum of disease and strategies for prediction and prevention," J. Rheumatol., vol. 51, pp. 337–349, 2024.
20. J. Demšar, "Statistical comparisons of classifiers over multiple data sets," JMLR, vol. 7, pp. 1–30, 2006.
21. D. Aletaha et al., "2010 RA classification criteria," Arthritis Rheum., vol. 62, pp. 2569–2581, 2010.
22. D. L. Scott et al., "Rheumatoid arthritis," The Lancet, vol. 376, pp. 1094–1108, 2010.
23. D. van der Heijde et al., "EULAR recommendations for RA management," Ann. Rheum. Dis., vol. 76, pp. 960–977, 2017.
24. A. Esteva et al., "Dermatologist-level classification of skin cancer with deep neural networks," Nature, vol. 542, pp. 115–118, 2017.

25. Z. Obermeyer and E. J. Emanuel, "Predicting the future — big data, ML, and clinical medicine," NEJM, vol. 375, pp. 1216–1219, 2016.
26. R. Caruana et al., "Intelligible models for healthcare: Predicting pneumonia risk and readmission," in KDD, 2015.
27. C. Molnar, Interpretable Machine Learning, 2nd ed., 2022.
28. M. T. Ribeiro et al., "Why should I trust you? Explaining classifier predictions," in KDD, 2016.
29. G. Varoquaux et al., "Assessing and tuning brain decoders: Cross-validation, caveats, and guidelines," NeuroImage, vol. 145, pp. 166–179, 2017.
30. T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning, 2nd ed. Springer, 2009.