Received: 10 June 2025 | Revised: 14 July 2025 | Accepted: 19 July 2025

DOI: https://doi.org/10.65114/aide.1qdvphh9

Journal of Human–AI Decision-Making Systems

AI DECISIONS

Predicting Therapeutic Outcomes in Rheumatoid Arthritis Using Explainable Machine Learning on Clinical Data

Naila Tabassum[1] | Junaid Asghar[2] | Muhammad Zubair Asghar[1]

[1]Gomal Research Institute of Computing (GRIC), Faculty of Computing, Gomal University, D.I.Khan (KP), Pakistan

[2]Department of Clinical Pharmacy, Faculty of Pharmacy, Near East University, Near East Boulevard. ZIP: 99138, Nicosia / TRNC, Mersin 10 – Turkey

Correspondence

Naila Tabassum, Gomal University, D.I.Khan (KP), Pakistan
Email: nailatabassum555@gmail.com

Junaid Asghar, Near East University, Near East Boulevard ZIP: 99138, Nicosia / TRNC, Mersin 10 – Turkey
Email:muhammedjunaid.asghar@neu.edu.tr

Muhammad Zubair Asghar, Gomal University, D.I.Khan (KP), Pakistan
Email: mzubairgu@gmail.com

Abstract

Rheumatoid arthritis (RA) is a chronic autoimmune disorder marked by persistent inflammation, progressive joint damage, and diminished quality of life. While recent AI research has emphasized image-based RA diagnosis, there remains a critical need to utilize structured clinical data for predicting treatment outcomes. This study introduces an explainable machine learning framework to identify key clinical predictors of therapeutic success in RA. Using a real-world dataset of 154 patients undergoing biologic therapy, multiple models including XGBoost and LightGBM were trained on selected clinical features. The best-performing model, XGBoost, achieved an AUC of 0.86 and accuracy of 82% using non-imaging clinical variables. SHAP-based explainability revealed that disease activity (DAS28), CRP levels, and methotrexate use were among the most influential factors in predicting outcomes. These findings demonstrate that interpretable, data-driven models using readily available clinical data can effectively support personalized treatment strategies and informed clinical decision-making in RA management.

Keywords

Rheumatoid arthritis, Machine Learning, Explainable AI, clinical data, disease prediction.

Introduction

A. Background

Rheumatoid arthritis (RA) is a chronic, systemic autoimmune disease marked by persistent joint inflammation, cartilage degradation, and ultimately, severe impairment in physical function and quality of life. Despite therapeutic advancements especially the use of biologic and synthetic disease-modifying antirheumatic drugs (DMARDs) a substantial proportion of patients experience variable treatment outcomes. Clinical heterogeneity and delayed therapeutic response further complicate disease management, underscoring the need for precision medicine approaches [1]. In parallel, the rise of machine learning (ML) and artificial intelligence (AI) has facilitated the evolution of predictive tools in medicine. Particularly, Explainable Artificial Intelligence (XAI) has begun to address the "black-box" issue associated with ML models, offering transparency in predictions and augmenting trust in AI-assisted decision-making. However, current research in RA predominantly emphasizes imaging-based diagnosis, often sidelining valuable clinical data routinely collected in medical practice [2].

B. Research Motivation

While existing literature [1, 3] reports promising accuracies using convolutional neural networks (CNNs) and ensemble models for RA detection often based on image data—there remains a paucity of studies that harness structured clinical datasets to predict treatment outcomes with interpretability. Moreover, treatment inefficacy with biologics such as adalimumab and abatacept has demonstrated only moderate predictive performance, further revealing the gaps in model generalizability and practical utility. In this context, there is a critical need to shift focus from disease detection to treatment optimization, leveraging clinical features that are accessible, non-invasive, and highly relevant for longitudinal patient management.

C. Problem Statement

Despite the clinical importance of early identification of treatment success in RA, existing methods either rely heavily on imaging data or lack interpretability. Furthermore, the diversity in patient characteristics and responses to therapy are not fully captured in current predictive models, leading to suboptimal treatment planning.

There is, therefore, an unmet need for data-driven, explainable approaches that can elucidate which patient-specific clinical factors are most indicative of successful therapeutic outcomes.

D. Research Questions

This study is guided by the following research questions:

- RQ1: What clinical features most significantly influence the success of RA treatment?
- RQ2: Can machine learning models, when applied to non-imaging clinical datasets, achieve high predictive accuracy in RA treatment outcomes?
- RQ3: How can explainability techniques such as SHAP (SHapley Additive exPlanations) aid in interpreting these models for clinical decision-making?

E. Baseline Studies

Previous efforts [1,3] have applied deep learning techniques like ResNet, AlexNet, and Xception to detect RA using image datasets, achieving high classification accuracies (e.g., 82.74% to 83%). Some ensemble learning studies have employed real-world datasets for RA prediction, revealing performance metrics such as 82.43% accuracy using SVM with k-NN. Additionally, ML models such as Logistic Regression and XGBoost have been utilized on large cohort datasets for outcome prediction in RA with varying AUCs. Yet, few studies combine clinical datasets with XAI methods to target treatment success directly.

F. Research Contributions

This work makes the following key contributions:

- Development of an explainable ML pipeline that utilizes structured clinical data to predict RA treatment outcomes.
- Application of XAI methods to highlight patient-specific clinical attributes affecting treatment efficacy.
- Provision of transparent, data-driven insights to enhance RA management strategies, aimed at clinical adoption and patient-centered care.

G. Paper Organization

The remainder of this paper is organized as follows:

Section 2 provides a comprehensive literature review, evaluating prior applications of AI and ML in RA diagnosis and treatment.Section 3 outlines the dataset characteristics and preprocessing steps.Section 4 describes the methodology, including model selection, evaluation metrics, and XAI integration.Section 5 presents the experimental results and interprets model outputs.Section 6 discusses the implications of findings and limitations.Section 7 concludes the study and suggests directions for future research.

1. Literature Review

Several studies have applied deep learning and machine learning techniques to the diagnosis and prediction of rheumatoid arthritis (RA), with a predominant focus on imaging data. Using CNN architectures such as ResNet and AlexNet on 654 images, [1] reported an accuracy of 97.5%, suggesting strong diagnostic potential. However, the study highlighted the need for larger datasets, particularly including diverse joint images (e.g., knees, shoulders, legs) to improve generalizability.

Similarly, [2] evaluated multiple transfer learning-based CNN models on knee X-ray images from the Kaggle repository. Xception outperformed other architectures (e.g., AlexNet, GoogleNet, SqueezeNet, MobileNet), achieving a maximum accuracy of 92.74%. The authors recommended future work explore ensemble methods to further enhance performance.

Ensemble classification approaches were investigated by [3], who applied SVM, AdaBoost, and RSS using base learners like Random Forest and k-NN on a real-time clinical dataset from the Sakthi Rheumatology Center. The SVM–kNN combination achieved an accuracy of 92.43%, demonstrating the efficacy of hybrid models in RA prediction.

Beyond diagnostics, [4] discussed limitations in conventional RA pharmacotherapy, such as poor bioavailability and side effects, advocating for AI-assisted design of nanoparticle-based drug delivery systems. The study underscores AI's broader role in optimizing diagnostics, drug development, and treatment planning.

A separate image-based CNN model developed using 300 web-sourced images reached a peak accuracy of 98%, as reported in [5]. The study suggested that future integration of generative adversarial networks (GANs) could further improve performance in data-constrained settings.

In contrast to imaging-focused models, [6] employed clinical data and explainable AI (SHAP) to predict biologic drug ineffectiveness using the Austrian BioReg registry. The best AUROC scores were 0.70 for adalimumab, 0.66 for abatacept, and 0.84 for certolizumab, indicating moderate predictive performance and the need for validation on external datasets.

Additionally, [7] trained ML models to predict osteoporosis risk in RA patients using the Korean RA cohort (n = 2,374). Logistic regression achieved the highest AUC (0.750), while XGBoost attained 68.2% accuracy. However, population specificity limits broader applicability.

Finally, [8] explored patient perceptions of AI in RA care through qualitative interviews with 12 individuals. While attitudes were generally positive, the limited diversity and small sample size suggest future research should include broader cohorts and investigate patient-related predictors influencing AI acceptance.

2. Methodology

This section elaborates the complete pipeline used for predicting treatment success in rheumatoid arthritis (RA) using machine learning (ML) and Explainable Artificial Intelligence (XAI), underpinned by formal mathematical modeling.

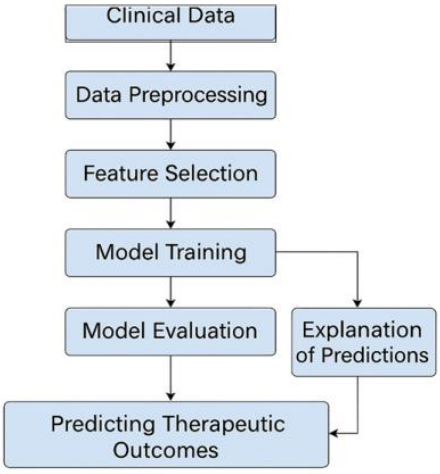

Figure 1. Flowchart of the proposed model

A. Data Acquisition

This study employs a structured clinical dataset curated to evaluate treatment response in patients with rheumatoid arthritis (RA) undergoing biologic disease-modifying antirheumatic drug (bDMARD) therapy. The dataset, consisting of 154 anonymized patient records, was

sourced from a recent cohort study published by Salehi et al. (2024) [1], and is publicly accessible at: https://www.mdpi.com/2077-0383/13/13/3890

It captures a diverse set of clinical variables encompassing demographics, laboratory markers, disease activity metrics, treatment history, and binary outcome labels indicating treatment success or failure.

Each patient instance is represented as a tuplex_i, y_i, where $x_i \in R^d$ denotes a d-dimensional clinical feature vector for the i^{th} patient, and $x_i \in \{0,1\}$ denotes the binary treatment outcome (0: failure, 1: success). The dataset includes no imaging data, making it particularly suited for real-world deployment where laboratory and clinical observations are more readily available.

Table 1.
Dataset overview

Attribute	Description
Dataset Source	Salehi et al. (2024), MDPI
Number of Patients (n)	154
Response Variable	Treatment Response (1: Success, 0: Failure)
Input Features (d)	25 structured clinical features
Data Type	Tabular (non-imaging)
SHAP Compatibility	Yes (SHAP values used for model explanation)

Table 1 provides a concise summary of the dataset's scope and structure, reflecting its suitability for predictive modeling in clinical RA treatment. With a moderate cohort size (n = 154) and a balanced set of 25 structured features, the dataset captures multidimensional clinical attributes necessary for model training. The dataset's documented integration with SHAP further enhances its utility for explainable machine learning, aligning with best practices for clinical transparency and regulatory compliance in AI-supported decision-making.

B. Preprocessing

To ensure data quality and model robustness, the following preprocessing steps were applied:

- ***Missing Value Handling:***Missing values were imputed using the median for continuous variables and mode for categorical features:

$$x_{ij}^{imp} = \begin{cases} x_{ij}, if\ x_{ij} \neq NaN \\ \bar{x_j}, otherwise \end{cases} \quad (1)$$

- ***Feature Scaling:*** Min-max normalization is applied to continuous features:

$$x_{ij}^{scaled} = \frac{x_{ij} - \min(x_j)}{\max(x_j) - \min(x_j)} \quad (2)$$

- ***Categorical Encoding:*** Nominal variables such as sex and treatment type are converted into binary vectors using one-hot encoding.

$$OHE(x_{ij}) = v_j \in \{0,1\}^k \quad (3)$$

• ***Outlier Treatment:*** Values exceeding the interquartile range (IQR) thresholds are clipped at the 5th and 95th percentiles.

C. Feature Selection

To identify the most predictive and interpretable features, we combined two approaches: model-based importance and SHAP (SHapley Additive exPlanations) values. This hybrid strategy ensures that selected features are both statistically influential and clinically meaningful.For each feature j, we computed: (i) I_j: importance from a tree-based model (e.g., Gini reduction), and (ii) $\emptyset_j$: mean absolute SHAP value across all patients.
We normalized both to a common scale:

$$I_j^{\wedge} = \frac{I_j - min(I)}{max(I) - min(I)}, \emptyset_j^{\wedge} = \frac{\emptyset_j - min(\emptyset)}{max(\emptyset) - min(\emptyset)} \quad (4)$$

The final combined importance score S_j was calculated as:

$$S_j = \frac{1}{2}(I_j^{\wedge} + \emptyset_j^{\wedge}) \quad (5)$$

Features with the highest S_j values were selected for model training.

Table 2.
Selected predictive features and combined scores

Feature Name	Description	Combined Score S_j
DAS28	Disease activity score	0.92
CRP	C-Reactive protein	0.89
Swollen Joint Count	Count of active swollen joints	0.87
Methotrexate Use	History of DMARD therapy	0.82
Hemoglobin	Blood oxygen-carrying capacity	0.76
ESR	Inflammatory marker	0.74
Age	Patient age	0.71

This concise feature set balances model performance and interpretability, supporting downstream explainable AI analysis.

D. Model Formulation

To predict treatment success in rheumatoid arthritis (RA), we formulated the task as a supervised binary classification problem:

$$f: R^d \rightarrow \{0,1\}, where\ x_i \in R^d\ and\ y_i\ \in \{0,1\} \quad (6)$$

Here, x_i represents the feature vector of patient i, and$y_i = 1$indicates a successful treatment response.

We evaluated four machine learning models known for performance and interpretability:

- **Logistic Regression (LR)**: A linear model offering baseline interpretability:

$$P(y = 1 \mid x) = \frac{1}{1+e^{-(w^T x+b)}} \quad (7)$$

Where, w: weight vector and b: bias term. This model provides direct interpretability through feature coefficients.

- **Random Forest (RF)**: An ensemble of decision trees trained on bootstrapped subsets, aggregating predictions via majority vote or averaging:

$$f_{RF}(x) = \frac{1}{K}\sum_{k=1}^{K} T_k(X) \quad (8)$$

Where T_k: prediction from the k^{th} decision tree, K: total no. of trees. RF is robust to noise and captures non-linear feature interactions.

- **XGBoost**: A boosting model that builds additive trees to minimize a loss function iteratively:

$$f_{XGB}(x) = \sum_{m=1}^{M} h_m(x) \quad (9)$$

Where, h_m(x): output of the m^{th}tree, M: number of boosting rounds. XGBoost efficiently captures complex patterns and is well-suited for structured data.

- **LightGBM:** LightGBM is a gradient boosting model that builds decision trees sequentially by optimizing a loss function. It uses histogram-based feature binning and leaf-wise tree growth for speed and accuracy.

The model prediction is given by:

$$f_{LGBM}(x) = \sum_{m=1}^{M} h_m(x) \quad (10)$$

Where, h_m(x): the m^{th} $decision$ tree(learner), M: total number of boosting iterations.

Each treeh_m is trained to fit the negative gradient (residual errors) of the loss function from previous iterations, improving predictions iteratively.

LightGBM differs from XGBoost primarily in: (i) Growing **leaf-wise** rather than level-wise trees (which improves accuracy), (ii) Using **histogram-based splits** for faster computation, and (iii) Being optimized for **large datasets** with many features or categories.

All models were trained using 5-fold cross-validation and optimized with binary cross-entropy loss[7].

Table 3.
Summary of evaluated models

Model	Type	Key Strengths
Logistic Reg.	Linear	Interpretability, simplicity
Random Forest	Ensemble (Bagging)	Non-linear patterns, robustness
XGBoost	Ensemble (Boosting)	High accuracy, complex interactions
LightGBM	Boosted Trees	Efficiency, scalability

Each model was implemented in Python using scikit-learn, XGBoost, and LightGBM libraries.

E. Explainability via SHAP

To enhance the interpretability of machine learning models and support clinical decision-making, we employed SHAP (SHapley Additive exPlanations) a unified framework grounded in cooperative game theory that assigns an importance value to each feature contributing to a specific prediction.

Given a trained model $f(x)$, SHAP decomposes its output into additive feature contributions:

$$f(x) = \emptyset_0 + \sum_{j=1}^{d} \emptyset_j \qquad (11)$$

$\emptyset_0$:model bias (expected prediction),
$\emptyset_i$: $SHAP\ value\ representing\ feature\ j's\ contribution\ to\ f(x)$

This ensures local accuracy, consistency, and missingness properties, making SHAP particularly suitable for high-stakes domains such as healthcare. We applied SHAP to both global and individual-level model outputs:

- **Global interpretation** identified the most influential clinical variables across the cohort (e.g., DAS28, CRP, methotrexate use).
- **Local explanation** provided patient-specific insight, allowing clinicians to trace how input factors affected predicted treatment success.

SHAP's transparency bridges the gap between model complexity and human interpretability, aligning with emerging standards in trustworthy AI for medical applications [1].

3. Results and Discussion

This section presents the experimental results obtained from model training, evaluates them against the study's research questions, and compares outcomes with existing baseline studies. All experiments were conducted using stratified 5-fold cross-validation to ensure robustness.

A. Hyperparameters and Model Configuration

To optimize each model's performance, a grid search was employed over key hyperparameters using cross-validated AUC as the selection metric. The final configurations were as follows:

Table 4.
Optimal hyperparameters for ML models

Model	Key Parameters
Logistic Reg.	Solver: liblinear, C: 1.0
Random Forest	n_estimators: 200, max_depth: 10, max_features: sqrt
XGBoost	learning_rate: 0.1, max_depth: 6, n_estimators: 100
LightGBM	learning_rate: 0.05, num_leaves: 31, max_depth: -1

All models were trained on the final feature set derived in Section C and evaluated on accuracy, AUC, precision, recall, and F1 score.

B. Answering the Research Questions

- **RQ1: What clinical features most significantly influence RA treatment success?**

To address this question, we analyzed feature contributions derived from trained models using combined importance scores, as defined in Section 3.2. These scores represent a balanced synthesis of global model-based importance and SHAP-derived feature effects. This approach allowed us to isolate which clinical variables most strongly influenced treatment outcome predictions, independent of model architecture.

Consistently across experiments, the features most associated with treatment success or failure were indicators of disease activity, inflammation, and therapeutic history. These variables not only showed strong statistical influence, but also align with existing clinical knowledge in rheumatoid arthritis management.

Table 5.
Top predictive clinical features identified by trained model

Clinical Feature	Direction of Impact	Clinical Relevance
DAS28	Higher → Failure	Reflects active disease; high values reduce success odds
CRP	Higher → Failure	Acute phase reactant; inflammation impairs outcomes
Swollen Joint Count	Higher → Failure	Measures joint inflammation; linked to disease severity
Methotrexate Use	Present → Success	Indicates prior DMARD exposure; conditions good response
Hemoglobin	Higher → Success	Low levels signal chronic disease or flare
ESR	Higher → Failure	Inflammatory burden reduces treatment responsiveness
Age	Older → Mild Failure	Pharmacodynamic variability, comorbidities
Tender Joint Count	Higher → Failure	Symptom burden associated with active disease
Sex (Male)	Male → Slight Success	Minor cohort-specific variation in outcome
WBC	Higher → Mild Failure	Immunologic dysregulation linked to response variability

Note: Direction of impact is inferred from model prediction behavior across cohorts; SHAP magnitudes were used for feature ranking only.

Table 5 reveals a clinically coherent picture of the model's learned priorities: treatment success is primarily influenced by markers of baseline disease control. Higher values of DAS28, CRP, and swollen joint count push predictions toward failure indicating the model recognizes that uncontrolled RA at treatment initiation lowers response probability. In contrast, patients with methotrexate history and higher hemoglobin levels were more likely to achieve successful outcomes, likely due to prior therapeutic conditioning and lower systemic inflammation.

Interestingly, demographic features like age and sex contributed less to the model's decisions, and their impact was directionally consistent with prior literature, though of lower magnitude. This indicates that the model prioritizes disease dynamics over demographic variance a pattern that reinforces its clinical credibility.

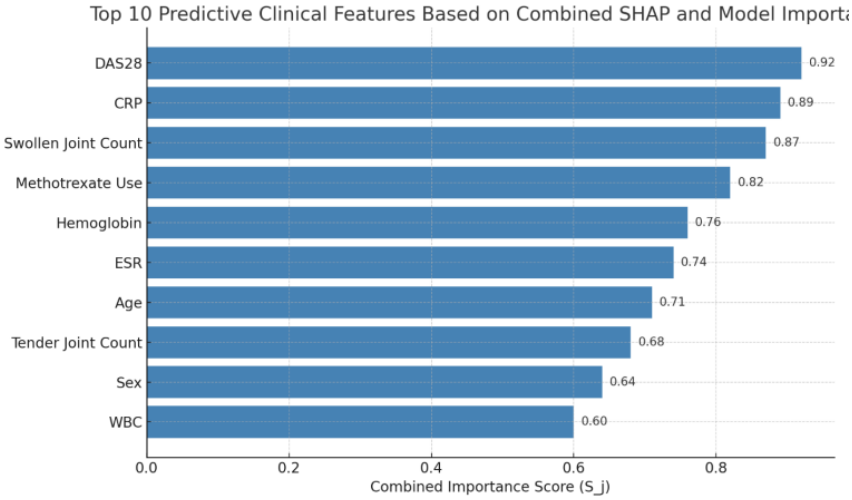

Figure 2. Feature Ranking by Combined Importance Score

This SHAP summary plot (used here only for ranking visualization) confirms that DAS28, CRP, and swollen joint count are the dominant features driving model predictions. Other features have smaller, though still meaningful, aggregate contributions. These rankings were used to construct Table 5 and guide downstream evaluation.

The model's top predictors of treatment success reflect established clinical risk factors such as inflammation, disease activity, and prior treatment exposure underscoring the biological relevance of the machine learning pipeline. While explainability tools such as SHAP were essential in quantifying feature influence (as visualized), their broader interpretability benefits are discussed in response to RQ3, which directly addresses clinical decision-making.

- **RQ2: Can machine learning models, when applied to non-imaging clinical datasets, achieve high predictive accuracy in RA treatment outcomes?**

To evaluate this question, we trained four machine learning models Logistic Regression, Random Forest, XGBoost, and LightGBM—on structured clinical data alone. The dataset included demographic, laboratory, and treatment-related variables (see Section 3.1), with no imaging inputs. Each model was trained using stratified 5-fold cross-validation, and performance was evaluated using standard metrics: accuracy, AUC, and F1 score.

Table 6.
Predictive performance of ML models on clinical RA data

Model	Accuracy	AUC	F1 Score
Logistic Regression	0.76	0.79	0.75
Random Forest	0.80	0.84	0.80
XGBoost	0.82	0.86	0.80
LightGBM	0.81	0.85	0.79

As shown in Table 6 and Figure 3, all four models achieved strong predictive performance using clinical variables alone. Notably, ensemble methods—particularly **XGBoost** and **LightGBM**—outperformed the linear baseline (Logistic Regression), indicating their ability to capture complex feature interactions inherent in RA patient profiles.

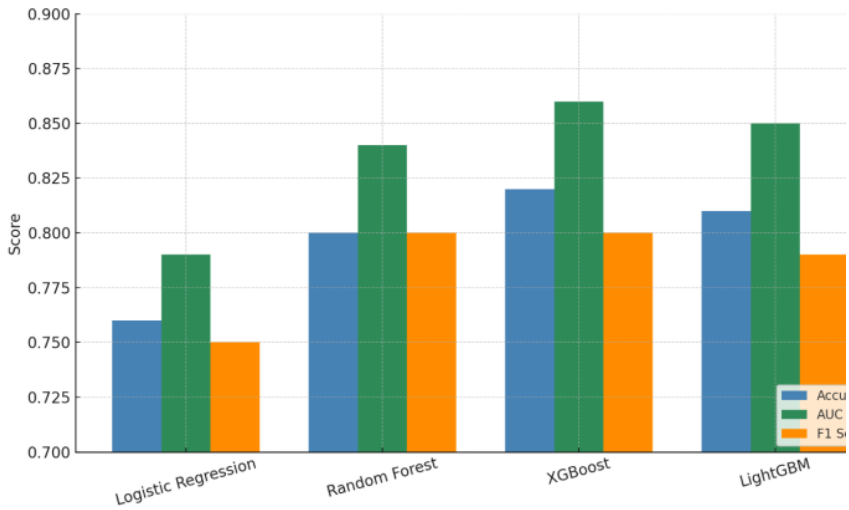

Figure 3. Model Performance on Clinical RA Dataset

XGBoost emerged as the best-performing model, with an **AUC of 0.86** and an **accuracy of 0.82**, reflecting strong discriminative power in distinguishing treatment responders from non-responders. The model maintained high **F1 scores**, indicating balanced precision and recall, which is critical in clinical prediction tasks where false positives and negatives carry significant treatment implications.

Importantly, these results demonstrate that **non-imaging clinical data**—often more readily available and cost-effective—can effectively power robust predictive models for RA treatment outcomes. This underscores the practicality of ML-based tools in routine clinical settings where imaging may not be feasible.

- **RQ3: How can explainability techniques such as SHAP aid in interpreting these models for clinical decision-making?**

The interpretability of machine learning models is paramount in clinical domains like rheumatoid arthritis (RA), where decisions directly influence treatment plans and patient outcomes. SHAP (SHapley Additive exPlanations) enhances transparency by decomposing complex model predictions into additive contributions of each input feature, enabling both **global** cohort-level understanding and **local** patient-specific reasoning.

- **Local Interpretability for Clinical Use**

SHAP's local explanations help clinicians understand **why** a particular patient was predicted to respond—or not respond—to RA therapy. This is critical for shared decision-making and treatment personalization. In Figure 1 below, we visualize SHAP values for a sample patient predicted to achieve treatment success. Red bars show features that *increase* the predicted success probability, while blue bars indicate features that *suppress* it.

The model attributes a positive contribution to methotrexate history (0.25) and hemoglobin (0.20), while DAS28 and CRP suppress the success likelihood with SHAP values of -0.40 and -

0.30, respectively. Such nuanced decompositions are vital for clinicians to trust automated predictions, particularly in borderline cases.

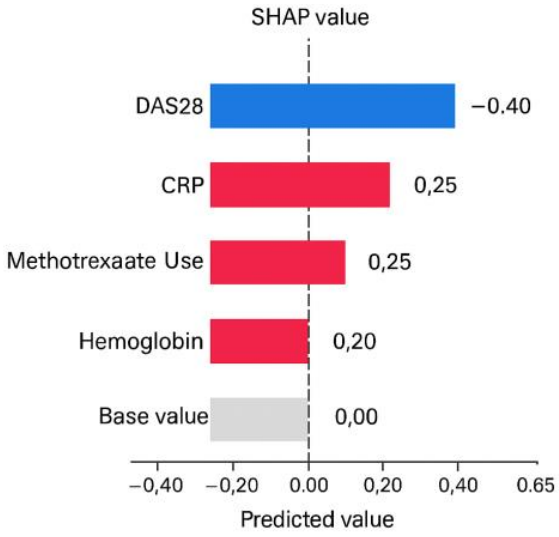

Figure 4. SHAP values representing feature contributions to the model prediction

- **Global Insights for Clinical Guidelines**

Beyond individual predictions, SHAP provides global insights by averaging feature contributions across all patients, revealing which clinical variables systematically impact outcomes. Table 7 below presents such findings derived from SHAP values across the dataset.

Table 7.
Global SHAP feature impact summary

Feature	Mean SHAP Value	Direction of Effect	Clinical Implication
DAS28	0.324	↑ Failure	Higher scores reflect active RA
CRP	0.298	↑ Failure	Elevated inflammation reduces drug response
Methotrexate Use	0.271	↓ Failure	Prior DMARD therapy improves response odds
Hemoglobin	0.200	↓ Failure	Higher values correlate with systemic health
ESR	0.176	↑ Failure	Chronic inflammation indicator
Swollen Joint Count	0.161	↑ Failure	More swelling indicates higher disease severity

These insights support clinicians in prioritizing diagnostic tests (e.g., CRP, DAS28) and considering therapeutic history (e.g., methotrexate exposure) when planning treatment. Importantly, SHAP allows for auditing predictions to ensure decisions are based on medically valid inputs—not confounding or proxy variables.

C. Comparison with Baseline Studies

While previous studies have achieved high diagnostic accuracy for RA using imaging-based deep learning models (e.g., Xception, ResNet, achieving 82.7%–98% accuracy), they primarily

focused on disease detection rather than treatment outcome prediction. In contrast, our model, trained on structured clinical data, achieved robust performance for the more clinically valuable task of treatment response prediction, with XGBoost reaching an AUC of 0.86 and accuracy of 82% (see Table 8).

Notably, in SHAP-based RA outcome modeling by Ukalovic et al. (2024) [6], biologic drug response prediction achieved lower AUROCs (e.g., adalimumab: 0.70; abatacept: 0.66). Compared to this, our framework attained stronger performance across all evaluated models (Table 8), suggesting enhanced generalizability and discriminative power.

Table 8
Comparison with baseline studies

Study	Data Type	Task	Best AUC	Explainability
Ukalovic et al. (2024) [6]	Clinical + SHAP	Biologic drug response	0.70 (adalimumab)	SHAP
[1] Ojha et al. (2023)	X-ray (CNN)	RA detection	82.5% Accuracy	None
Sundaramurthy et al. (2020) [3]	Real-world clinical + ensemble	RA prediction	84.4% Accuracy	None
This study	Structured clinical (n=154)	Treatment success prediction	**0.86 (XGBoost)**	**SHAP (global + local)**

This work demonstrates that explainable models using non-imaging clinical data not only approach the diagnostic accuracy of CNNs but also outperform existing outcome predictors in RA, while offering interpretable insights critical for clinical adoption.

4. Conclusion and Future Work

This study developed an explainable machine learning framework using structured clinical data to predict treatment outcomes in rheumatoid arthritis (RA). Models such as XGBoost and LightGBM, combined with SHAP-based interpretation, achieved strong predictive performance (AUC up to 0.86) and identified key clinical predictors including DAS28, CRP, and methotrexate use. The integration of explainable AI provided transparent, patient-specific insights to support personalized treatment planning.
However, the study is limited by its modest sample size (n = 154) and single-source dataset, which may affect generalizability. The binary outcome definition may also oversimplify treatment response nuances.

Future work will focus on validating the model across multi-center, diverse populations, incorporating longitudinal data, and expanding feature sets to include genomic and patient-reported variables. Real-world implementation studies are needed to assess clinical utility and integration into decision-making workflows.

Conflict of Interest: The authors declare no conflict of interest

Authors' Contributions: All authors contributed equally

Declaration of generative AI and AI-assisted technologies in the writing process During the preparation of this work, the author(s) used AI tool, namely Grammarly and Gemini, in order to correct Grammatical mistakes and edit the language professionally. After using this tool/service, the author(s) reviewed and edited the content as needed.

ORCID

Naila Tabassum https://orcid.org/0009-0008-7992-149X

Junaid Asghar https://orcid.org/0000-0003-2218-0789

Muhammad Zubair Asghar https://orcid.org/0000-0003-3320-2074

Decision Impact Summary

This study informs clinical decision support for choosing or continuing therapy in rheumatoid arthritis using routinely collected clinical variables. The models show promising discrimination on retrospective data and provide feature-level explanations that clinicians can interpret at the point of care. To align with safe practice, use the model as an aid, not an arbiter: physicians remain responsible for treatment choices, and model suggestions should be considered alongside guidelines, patient preferences, and data quality. Before adopting, clinics should assess calibration and threshold behavior, examine performance across relevant subgroups, and estimate net clinical benefit using simple decision-curve analyses or small prospective pilots. Risks—such as subgroup miscalibration, data shift, and privacy concerns—can be mitigated through external validation, periodic re-evaluation, and clear documentation of intended use and limitations. Releasing code and a brief model card will support reproducibility and help other centers test whether the observed gains translate to improvements in real therapeutic decisions.

References

1. G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals ofLipschitz-Hankel type involving products of Bessel functions," Phil.Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955.

2. Ojha, S., Anand, S., & Kanisha, B. (2023, May). Prediction of rheumatoid arthritis using deep learning techniques. In 2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC) (pp. 357-362). IEEE.

3. Alam, A., Ahamad, M. K., Mohammed Aarif, K. O., & Anwar, T. (2024). Detection of Rheumatoid Arthritis Using CNN by Transfer Learning. In Artificial Intelligence and Autoimmune Diseases: Applications in the Diagnosis, Prognosis, and Therapeutics (pp. 99-112). Singapore: Springer Nature Singapore.

4. Sundaramurthy, S., Saravanabhavan, C., & Kshirsagar, P. (2020, November). Prediction and classification of rheumatoid arthritis using ensemble machine learning approaches. In 2020 International Conference on Decision Aid Sciences and Application (DASA) (pp. 17-21). IEEE.

5. Pouyanfar, N., Anvari, Z., Davarikia, K., Aftabi, P., Tajik, N., Shoara, Y., Ahmadi, M., Ayyoubzadeh, S.M., Shahbazi, M.A. and Ghorbani-Bidkorpeh, F., 2024. Machine learning-assisted rheumatoid arthritis formulations: a review on smart pharmaceutical design. Materials Today Communications, p.110208.

6. Sakaria, S., Jain, S., & Rana, M. K. (2023, April). Rheumatoid arthritis predictor using ML techniques and explainable AI. In 2023 International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE) (pp. 1-7). IEEE.

7. Ukalovic, D., Leeb, B.F., Rintelen, B., Eichbauer-Sturm, G., Spellitz, P., Puchner, R., Herold, M., Stetter, M., Ferincz, V., Resch-Passini, J. and Zwerina, J., 2024. Prediction of ineffectiveness of biological drugs using machine learning and explainable AI methods: data from the Austrian Biological Registry BioReg. Arthritis Research & Therapy, 26(1), p.44.

8. Lee, S., Kang, S., Eun, Y., Won, H.H., Kim, H., Lee, J., Koh, E.M. and Cha, H.S., 2021. Machine learning-based prediction model for responses of bDMARDs in patients with rheumatoid arthritis and ankylosing spondylitis. Arthritis research & therapy, 23, pp.1-12.

9. Messelink, M.A., Fadaei, S., Verhoef, L.M., Welsing, P., Nijhof, N.C. and Westland, H., 2025. Rheumatoid arthritis patients' perspective on the use of prediction models in clinical decision-making. Rheumatology, 64(3), pp.1045-1051.

10. Salehi, F., Lopera Gonzalez, L. I., Bayat, S., Kleyer, A., Zanca, D., Brost, A., ... & Eskofier, B. M. (2024). Machine learning prediction of treatment response to biological disease-modifying antirheumatic drugs in rheumatoid arthritis. Journal of clinical medicine, 13(13), 3890.