Düzce University Journal of Science and Technology
Düzce Üniversitesi Bilim ve Teknoloji Dergisi

14 (2), 2026, 695–704.
https://doi.org/10.29130/dubited.1768590

e-ISSN: 2148-2446
https://dergipark.org.tr/en/pub/dubited

Research Article

Study on Rheumatoid Arthritis Prediction Model With Machine Learning and Ensemble Learning Methods, and Demographic, Clinical, Laboratory Characteristics of These Patients

Kemal Üreten[1*], Yüksel Maraş[2], Ebru Atalar[2], Kevser Orhan[2]

[1] Yozgat Bozok University, Faculty of Medicine, Department of Internal Medicine, Yozgat, Türkiye.
ror.org/04qvdf239

[2] Ankara Bilkent City Hospital, Rheumatology Clinic, Ankara, Türkiye.
ror.org/033fqnp11

*Corresponding author:
kemalureten@yahoo.com

Received: 19.08.2025
Revision: 29.12.2025
Accepted: 29.03.2026
Online: 24.07.2026

Abstract: The aim of this study is to develop a diagnostic and predictive model for rheumatoid arthritis (RA) with machine learning and ensemble learning methods using data from patients who applied to the rheumatology outpatient clinic with complaints of pain in the hand joints and to review the demographic, clinical and laboratory results of these RA patients. This study was conducted on demographic properties, clinical findings, and laboratory test results of a total of 421 patients, 260 with RA and 161 non-RA. Machine learning and ensemble learning algorithms were used for prediction. In this study, ensemble learning algorithms showed better performance than machine learning algorithms. With the Random Forest Classifier model 88% accuracy (95% confidence interval (CI), 81% - 94%), 94% sensitivity (95% CI, 88% - 99%), 78% specificity (95% CI, 69% - 86%), 87% positive predictive value (PPV) (95% CI, 79% - 94%), 90% F1 score (95% CI, 84% - 96%), and 0.92 AUC (area under the curve) results were obtained. All three developed ensemble learning models showed similar performance in statistical metrics. These developed models can be improved with new multicenter studies with more patients and can be used in daily practice.

Key words: *Rheumatoid arthritis, Machine learning, Ensemble learning, Logistic regression, Support vector machines*

I. INTRODUCTION

Rheumatoid arthritis (RA) is an inflammatory arthritis that begins insidiously and affects mainly the small joints of the hands and feet. In later stages, it may also affect other synovial joints such as the elbow, shoulder, knee, and hip. It affects 0.5-1% of the population and is 3-5 times more common in women than in men (Almutairi et al., 2021; Radu & Bungau, 2021).

RA is a multifactorial disease caused by genetic, environmental, and various unknown factors. A family history of rheumatoid arthritis (RA) is a strong risk factor for developing RA, with first-degree relatives of patients with RA having a two to fourfold increased risk of developing the disease. Smoking, air pollution, dietary habits, and infectious agents influence disease incidence and progression (Deane & Holers, 2019; Otón & Carmona, 2019; Romão & Fonseca, 2021).

RA can be found together with many autoimmune diseases, such as thyroiditis, vitiligo, Sjogren syndrome, and diabetes mellitus (DM). Autoimmune thyroiditis (AT) is the most common human autoimmune disorder (Conigliaro et al., 2020), and AT patients often have another autoimmune disorder, such as RA, pernicious anemia, systemic lupus erythematosus, vitiligo (Fallahi et al., 2016).

Depending on the presence or absence of antibodies to rheumatoid factor (RF) and anti-cyclic citrullinated protein/peptide antibody (anti-CCP), RA can be divided into two types: seropositive and seronegative RA. The results of both tests are conclusive. A positive RF or anti-CCP test result increases overall diagnostic sensitivity. When both tests are positive together, the specificity increases (Abdelhafiz et al., 2023; Jindal et al., 2021). In addition to anemia, erythrocyte sedimentation rate (ESR) and serum C-reactive protein (CRP) levels are elevated in RA patients due to activation of chronic inflammation (Mun et al., 2021).

Machine learning (ML) is a branch of computer science and artificial intelligence that allows computers to learn from existing data using various algorithms without being explicitly programmed. The goal is to make predictions about the future based on existing data. Nowadays, with the rapid development of technology, producing, storing and accessing data is becoming much easier. This data is valuable as long as it is converted into meaningful information. Machine learning algorithms are successfully used in classification, regression, clustering tasks. Decision trees, linear regression, logistic regression, naive bayes, k means clustering, support vector machines, random forest algorithms and artificial neural networks are well-known and frequently used machine learning algorithms. Machine learning methods have been successfully used in areas such as fraud detection, credit scoring, image recognition, healthcare, and natural language

Cite as: Üreten, K., Maraş, Y., Atalar, E., & Orhan, K. (2026). Study on rheumatoid arthritis prediction model with machine learning and ensemble learning methods, and demographic, clinical, laboratory characteristics of these patients. *Düzce University Journal of Science and Technology, 14*(3), 695–704.
https://www.doi.org/10.29130/dubited.1768590

This is an open access paper distributed under the terms and conditions of the Creative Commons Attribution-NonCommercial 4.0 International License.

processing. (Dong et al., 2020; Janiesch et al., 2021; Mahesh, 2018; Mutasa et al., 2020; Sagi & Rokach, 2018; Sarker, 2021).

Ensemble learning is a machine learning technique that allows building a model with multiple learners instead of training the model with a single learner. Ensemble learning is a powerful technique to improve the performance of machine learning models by combining multiple individual models. By taking advantage of the diversity and collective intelligence of the ensemble, it can provide more accurate and reliable predictions for complex tasks. It helps to reduce overfitting. Ensemble learning can be applied to both classification and regression tasks. Ensemble learning algorithms are classified into 3 groups; bagging, boosting, and stacking (Sagi & Rokach, 2018). Bagging is an ensemble learning algorithm that combines multiple base learners to create an accurate and robust model. It is an effective technique used to solve classification and regression problems in machine learning. Random forest algorithm is an example of bagging (Mahesh, 2018). Unlike bagging, where the base learners are trained independently, boosting trains the base learners sequentially. Initially, equal weights are assigned to all observations, creating multiple models that each correct the errors of the previous model. The strengths of each algorithm are exploited and its weaknesses are compensated. AdaBoost, Gradient Boosting, and XGBoost are examples of boosting algorithms (Azmi & Baliga, 2020; Sagi & Rokach, 2018). The stacking method is ideal for solving complex problems and is created by combining different types of models. Each model in the ensemble complements each other and the output of the next model is used as the input of the previous model. In this way, the models can interact with each other to produce more accurate results. Stacking ensemble learning is especially useful when working with large datasets and can be used to reduce noise and obtain accurate results.

Some studies have been conducted on the diagnosis and prediction of diseases in the medical field using machine learning methods. These include studies on the diagnosis of coronary heart disease, COVID-19 diagnosis, acute rheumatic fever in children, diabetes, and chronic renal failure (Ali et al., 2021; Dritsas & Trigka, 2022; Emre et al., 2019; Faruque et al., 2019; Muhammad et al., 2021). Most studies on RA in recent years have focused on disease management and relapses; there are few studies on the diagnosis and prediction of the RA. These studies have not investigated which features are important in predicting the disease. RF and anti-CCP are important serological tests in RA diagnosis; anti-CCP, in particular, is not readily available in every laboratory, and the success of models developed without RF and anti-CCP results in the dataset has not been evaluated (Mendoza-Pinto et al., 2024).

Sundaramurthy et al. (2020) conducted a study on RA prediction and classification. Their dataset included 250 healthy and 750 patients with RA. They used the patients' age, gender, RF, anti-CCP, morning stiffness, symmetric joint inflammation, rest pain, ESR, body pain and swelling characteristics. They used an ensemble machine learning algorithms and achieved 94% accuracy, 90.80% positive predictive value (PPV) and 94.70% ROC results with support vector machines (SVM) (Sundaramurthy et al., 2020).

In their study to predict the prognosis of RA, Wu et al. (2020) studied on 3081 complicated and 405 uncomplicated RA patients. They evaluated the patients' demographic data, laboratory results, presence of heart disease, hepatitis and end-stage renal disease, and medications taken by the patients. They found that decision tree-based techniques provided more accurate results (Wu et al., 2020).

Matsuo et al. (2022) in their study used 73 ultrasound features and 54 blood test results to predict RA recurrence during 2-year follow-up. They had a total of 210 patients, of whom 60 patients relapsed and 150 patients were in remission during this period. For relapse prediction, they used the logistic regression, random forest, and Extreme Gradient Boosting (XGBoost) algorithms. Using the XGBoost classifier, they obtained a result of 0.747 for the area under curve (AUC) (Matsuo et al., 2022).

The aim of this study is to develop a RA diagnostic and predictive method using the demographic information, medical history, clinical findings and laboratory blood tests of patients who applied to health centers with complaints of pain in the hand joints and to compare the performances of machine learning and ensemble learning algorithms for this task, and to determine which features are important in developing the RA predictive model. Additionally, this study aims to evaluate the performance of models developed without anti-CCP and RF serological test results, which are not readily available in every center.

II. MATERIAL AND METHOD

A. Dataset

This retrospective study was conducted using the medical records of patients who were examined at the rheumatology outpatient clinic of Ankara Bilkent City Hospital between January and July 2023 with complaints of pain in the hand joints. Ethics committee permission was obtained from the local ethics committee. The results of the patients' age, gender, body mass index (BMI), family history of RA, hypertension (HT), diabetes mellitus (DM), vitiligo, thyroiditis, smoking, morning stiffness, swelling-tenderness of hands, hemoglobin, white blood cells (WBC), neutrophils, lymphocytes, platelet count (plt), erythrocyte sedimentation rate (ESR), C-reactive protein (CRP), rheumatoid factor (RF), anti-cyclic citrullinated protein/peptide antibody (anti-CCP) levels were used. Patients were classified as RA and

non-RA patients group. The non-RA group consisted of patients diagnosed with nonspecific arthralgia, osteoarthritis, myalgia, tendonitis and Sjogren's syndrome. There were 260 patients in the RA group and 161 patients in the non-RA group, 335 of the patients were women and 86 were men, the average age of the RA patient group was 57.1, and the average age of the non-RA group was 50.5.

This dataset, which included patients' clinical and demographic characteristics, family history, and serological test results, was named dataset-1. RF and anti-CCP, which are important in RA diagnosis and not available at every center, were removed from the dataset, and a new dataset was created (dataset-2). The aim of this was to test the performance of the models with patients' other characteristics without the RF and anti-CCP serological test results.

B. Data Preprocessing

Gender, family history of RA, HT, DM, vitiligo, thyroiditis, smoking, morning stiffness, swelling-tenderness of hands values were categorical variables, hemoglobin, WBC, neutrophils, lymphocytes, plt, ESR, CRP, RF and anti-CCP levels were numerical variables. The data set was checked for missing data, there were no missing data in categorical variables. 19 patients' WBC and plt, 18 patients' Hb, neutrophils, lymphocytes and CRP, 17 patients' ESR, 70 patients' BMI, 55 patients' RF and 113 patients' anti-CCP results were missing. To prevent data leakage, an 80:20 train-test split was implemented at the beginning of the study; test data was not used during training, but was used to evaluate the final performance of the developed models. Missing data in continuous variables was completed using the Multiple Imputation by Chained Equations (MICE) method, widely accepted in machine learning studies (Azur et al., 2011; Luo, 2022). The *IterativeImputer* method of the sklearn library was used for this. Iterative imputation was fitted on the training data (*estimator = BayesianRidge(), max_iter=20*), and the test data was only transformed, thus preventing data leakage and eliminating data deletion, thereby preventing potential biases and data loss. Standardization was applied after multiple imputation. Hyperparameter optimization of the algorithms was carried out by drawing on similar studies conducted previously and through trial and error. A 5-fold stratified cross-validation was conducted to evaluate the stability of the developed models, and mean ± standard deviation was reported for accuracy, sensitivity, specificity, and AUC performance metrics, with p-values calculated. The final evaluation was performed using test data that was set aside at the beginning of the study and never used during training. Whether there was a significant difference in patients RA and non-RA in terms of categorical variables was evaluated with the chi-square test, and p values were calculated. The RF cut off level for our hospital laboratory is 15 IU/mL and the anti-CCP cut off level is 5U/mL. Using these cut-off levels, RF and anti-CCP variables were converted into categorical data and p values were calculated with chi-square test. Categorical variables are ordinal in structure and were numerically encoded during the data collection phase (e.g., 0/1 or 1/2). Therefore, no additional encoding was applied.

This study used the Python 3.9 programming language, imported the pandas, numpy, matplotlib, seaborn, and sklearn libraries, and utilized the sklearn library for machine learning algorithms and statistical calculations. The *IterativeImputer* algorithm from the sklearn.impute library was used for imputation, and the *feature_importances_* function from RandomForestClassifier was used to identify important features for classification.

C. Feature Importance

Feature quality is important in building a machine learning model. The feature importance indicates how much each feature contributes to the prediction of the model. Some features are more important to the model, while others are less important. Essentially, it determines the degree of usefulness of a particular variable to an existing model and prediction. Feature importance scores can be calculated in decision trees, random forests, and linear model algorithms and features can be ranked according to the calculated scores. In this study, feature importance scores were calculated and graphed using the built-in *feature_importances_* function of the RandomForest classifier (Carvalho et al., 2019; Castiglioni et al., 2021).

D. Statistical Analysis

The performance of the model was evaluated with performance metrics calculated from the confusion matrix obtained with the test data. The confusion matrix is used to compare the predicted values with the actual values. Accuracy, sensitivity, specificity, positive predictive value (PPV), F1 score and Cohen's kappa are calculated using the numbers from TP (True Positive), FP (False Positive), TN (True Negative) and FN (False Negative). AUC (area under the curve) is obtained by calculating the area under the ROC (receiver operating characteristic) curve, which is used to determine the discrimination power of a model.

$$\boldsymbol{accuracy} = \frac{\boldsymbol{TP} + \boldsymbol{TN}}{\boldsymbol{TP} + \boldsymbol{TN} + \boldsymbol{FP} + \boldsymbol{FN}}$$

$$\boldsymbol{sensitivity}\ (\boldsymbol{recall}) = \frac{\boldsymbol{TP}}{\boldsymbol{TP} + \boldsymbol{FN}}$$

$$\boldsymbol{specificity} = \frac{\boldsymbol{TN}}{\boldsymbol{TN} + \boldsymbol{FP}}$$

$$\textbf{positive predictive value}(\textbf{PPV}) = \frac{\boldsymbol{TP}}{\boldsymbol{TP + FP}}$$

$$\textbf{F1 score} = \frac{\boldsymbol{2 * (precision * recall)}}{\boldsymbol{(precision + recall)}} = \frac{\boldsymbol{2TP}}{\boldsymbol{2TP + FP + FN}}$$

$$\textbf{chi - square test} = \frac{\boldsymbol{(Observed - Expected)2}}{\boldsymbol{Expected}}$$

$$\textbf{Odds ratio} = = \frac{a/c}{b/d} = \frac{a*d}{b*c}$$

III. RESULTS AND DISCUSSIONS

This study utilized decision trees, support vector machines, k-nearest neighbors, and logistic regression machine learning algorithms. Ensemble learning models were developed using random forest classifiers for bagging, adaboost classifiers for boosting, and random forest and k-nearest neighbors classifiers for stacking. The parameters of the ensemble learning models set in this study are shown in Table 1.

Table 1. The base estimators and hyperparameters of the ensemble learning models.

Model	**base_estimator, classifier, parameters**
Bagging	Base_estimator = RandomForestClassifier(Class_weight = 'balanced', criterion = 'entropy', max_depth =11, min_samples_leaf = 6, min_samples_split = 6)
Boosting	AdaBoostClassifier (DecisionTreeClassifier(max_depth = 5), (n_estimators = 400))
Stacking	base_estimator =[(RandomForestClassifier(n_estimators=200), (KNeigborsClassifier(n_neigbors=4)], meta_estimator = LogisticRegression

In this study, ensemble learning algorithms performed well, and similar results were obtained in statistical metrics. With the Random Forest classifier, 88% accuracy (95% CI, 81-94%), 94% sensitivity (95% CI, 88-99%), 78% specificity (95% CI, 69-86%), 87% positive predictive value (95% CI, 79-94%), 90% F1 score (95% CI, 84-96%), and 0.92 AUC (area under the curve) were obtained. Table 2 shows the final test results for accuracy, sensitivity, specificity, PPV, F1 score, and AUC scores of the models. Table 3 shows the means and standard deviations of the accuracy, sensitivity, specificity, and AUC results obtained from the 5-fold stratified cross-validation. Figure 1 shows the confusion matrix and ROC curve obtained by the Random forest classifier. Figure 2 shows the ranking of features obtained with the Random forest classifier's *feature_importances_* function.

Table 2. Performance metrics obtained with the machine learning and ensemble learning models (dataset-1).

		Accuracy	Sensitivity	Specificity	PPV	F1 score	AUC
Machine learning Algorithms	Logistic regression	0.85	0.90	0.75	0.85	0.87	0.93
	Support vector machines	0.78	0.84	0.69	0.81	0.83	0.88
	Decision trees	0.85	0.96	0.67	0.82	0.88	0.87
	K neigbour classifier	0.75	0.82	0.63	0.78	0.80	0.76
Ensemble learning Algorithms	Random forest	0.88	0.94	0.78	0.87	0.90	0.92
	AdaBoost classifier	0.88	0.96	0.75	0.87	0.90	0.93
	Stacking classifier	0.88	0.98	0.72	0.85	0.91	0.95

Table 3. Performance metrics obtained from 5-fold stratified cross-validation (dataset-1).

		Accuracy (mean±Sd)	Sensitivity (mean±Sd)	Specificity (mean±Sd)	AUC (mean±Sd)
Machine learning Algorithms	Logistic regression	0.72±0.05	0.76±0.02	0.67±0.10	0.80±0.06
	Support vector machines	0.73±0.04	0.75±0.09	0.70±0.12	0.80±0.06
	Decision trees	0.73±0.02	0.82±0.05	0.59±0.09	0.77±0.03
	K neigbour classifier	0.62±0.01	0.78±0.03	0.35±0.04	0.61±0.01
Ensemble learning Algorithms	Random forest	0.79±0.05	0.87±0.08	0.66±0.11	0.88±0.04
	AdaBoost classifier	0.77±0.02	0.82±0.04	0.68±0.0.07	0.84±0.01
	Stacking classifier	0.79±0.01	0.91±0.03	0.57±0.03	0.84±0.02

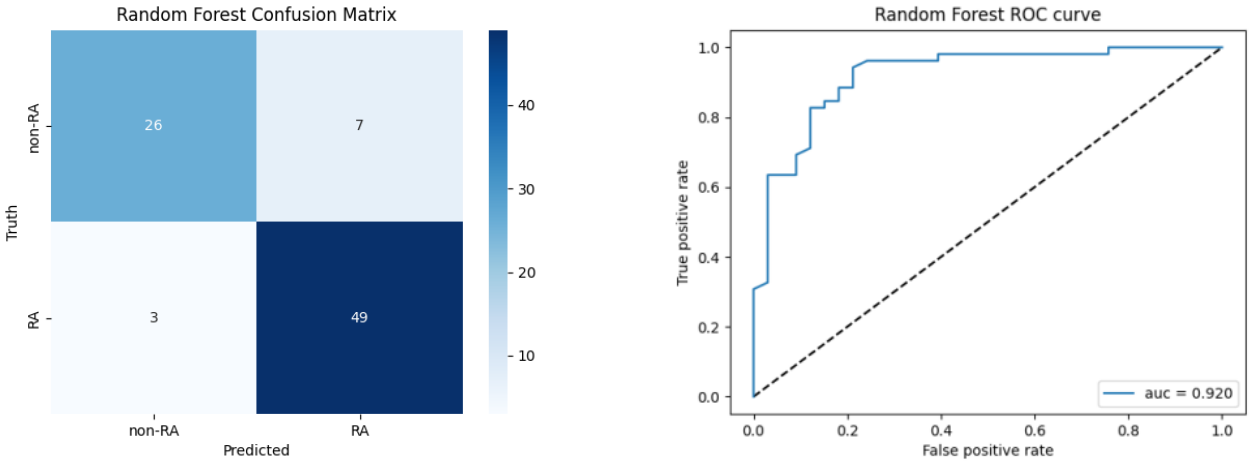


Figure 1. Confusion matrix and ROC curve obtained with the Random forest classifier.

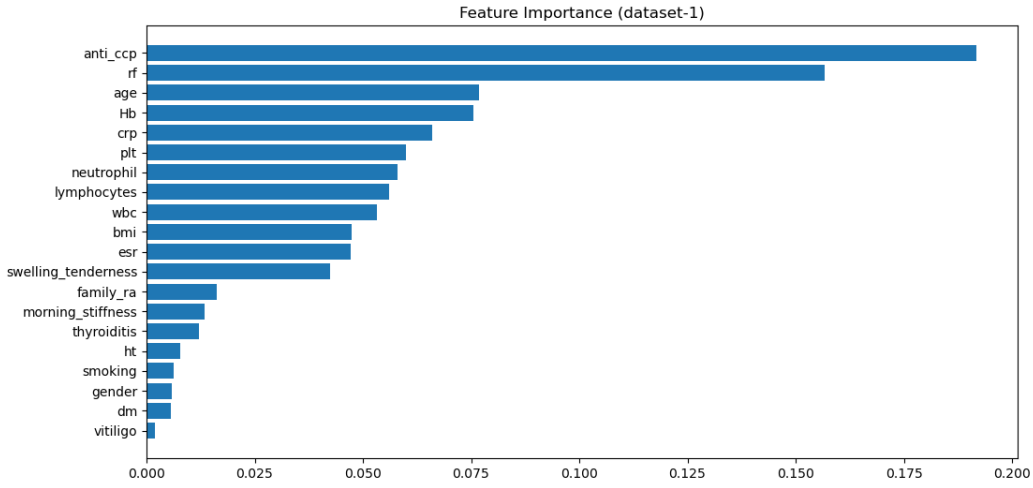


Figure 2. Ranking of features with feature importance function (dataset-1).

The final test results of the models obtained with dataset-2, which did not include RF and anti-CCP serological test results, are shown in Table 4, and the test results obtained with 5-fold stratified cross-validation are shown in Table 5. Figure 3 shows the ranking of features obtained with dataset-2 using the Random forest classifier's *feature_importances_* function. Table 6 shows the p-values obtained by comparing categorical variables in RA and non-RA patient groups with the chi-square test, and the odds ratios of categorical variables.

Table 4. Performance metrics obtained with the machine learning and ensemble learning models (dataset-2).

		Accuracy	Sensitivity	Specificity	PPV	F1 score	AUC
Machine learning Algorithms	Logistic regression	0.72	0.77	0.64	0.77	0.77	0.75
	Support vector machines	0.61	0.87	0.21	0.63	0.73	0.68
	Decision trees	0.65	0.69	0.59	0.72	0.71	0.69
	K neigbour classifier	0.69	0.83	0.48	0.78	0.72	0.67
Ensemble learning Algorithms	Random forest	0.74	0.88	0.51	0.74	0.80	0.73
	AdaBoost classifier	0.72	0.92	0.39	0.70	0.75	0.95
	Stacking classifier	0.75	0.92	0.48	0.74	0.82	0.73

Table 5. Performance metrics obtained from 5-fold stratified cross-validation (dataset-2).

		Accuracy (mean±Sd)	Sensitivity (mean±Sd)	Specificity (mean±Sd)	AUC (mean±Sd)
Machine learning Algorithms	Logistic regression	0.73±0.03	0.84±0.07	0.55±0.07	0.75±0.02
	Support vector machines	0.66±0.03	0.88±0.05	0.31±0.08	0.65±0.05
	Decision trees	0.63±0.05	0.65±0.10	0.60±0.09	0.68±0.05
	K neigbour classifier	0.57±0.02	0.77±0.03	0.23±0.04	0.54±0.04
Ensemble learning Algorithms	Random forest	0.72±0.05	0.87±0.06	0.48±0.09	0.75±0.02
	AdaBoost classifier	0.68±0.04	0.88±0.05	0.36±0.09	0.72±0.03
	Stacking classifier	0.68±0.08	0.88±0.07	0.34±0.11	0.72±0.07

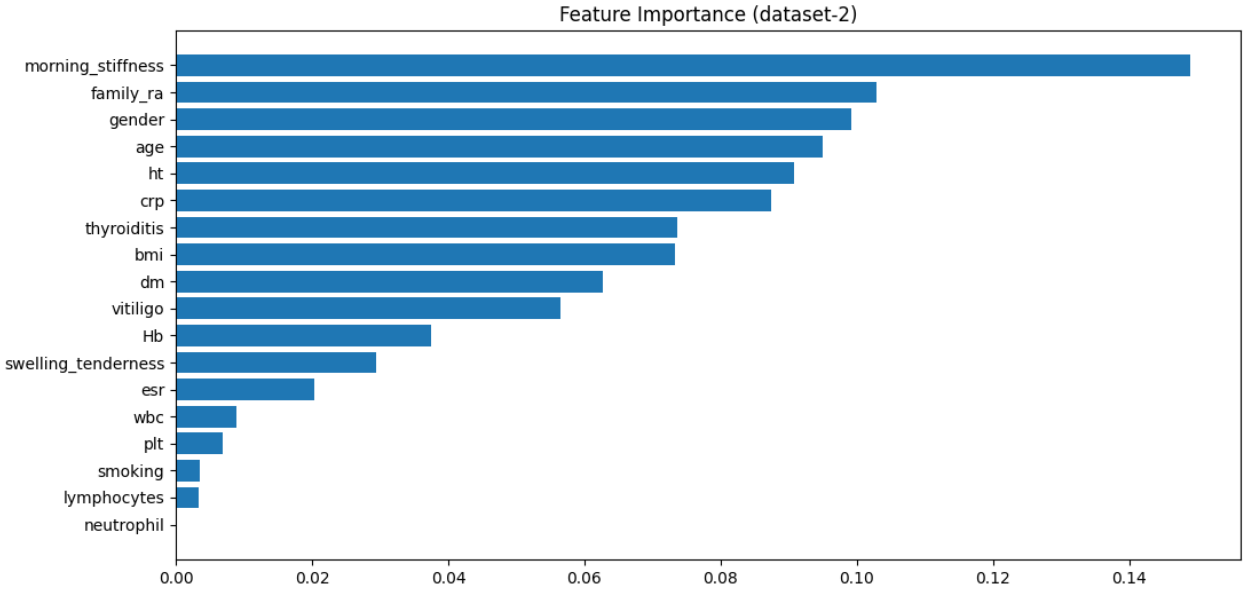


Figure 3. Ranking of features with feature importance function (dataset-2).

Table 6. p values obtained by comparing categorical variables in RA and non-RA patient groups with chi-square test, and Odds ratios of categorical variables.

	p value	Odds ratio (OR)	95% Confidence Interval (CI)
Gender	0.47	1.20	0.73 - 1.97
Family history	< 0.001	2.98	1.70 - 5.23
Hypertension	0.003	2.06	1.27 - 3.33
Diabetes mellitus	0.52	1.21	0.67 - 2.16
Vitiligo	0.16	4.43	0.54 - 36.32
Thyroiditis	< 0.001	3.73	1.77 - 7.83
Smoking	0.53	1.16	0.72 - 1.88
Morning stiffness	0.001	2.08	1.32 - 3.28
Swelling-tenderness	< 0.001	3.82	2.41 - 6.06
RF (cut off = 15)	< 0.001	12.3	7.10 - 21.58
anti-CCP (cut off = 5)	< 0.001	38.4	18.07 - 81.94

The diagnosis of RA is made by evaluating together the patient's demographic data, medical history, examination findings, laboratory results and radiological findings. With early diagnosis and appropriate treatment, the development of deformity can be prevented and the patient's quality of life can be improved. In this study, which was conducted using machine learning and ensemble learning methods to differentiate patients presenting to the rheumatology outpatient clinic with complaints of pain in the hand joints and to develop a prediction model for RA, the demographic characteristics, medical history, clinical findings, and laboratory test results were used.

Firstly, the characteristics of the RA patient group in this study were compared with the general literature knowledge of patients with RA. In the general population, RA is 3-5 times more common in women than in men. In this study, RA was approximately four times more common in women, but gender was not a statistically significant risk factor for RA (Odds ratio (OR = 1.20, p = 0.47)). RA is 2-4 times more common in first-degree relatives of RA patients and is a strong risk factor for developing RA. (Frisell et al., 2016). In our study group, 71 of 260 RA patients had a first-degree relative with RA (OR = 2.98, p < 0.001), indicating that a family history of RA is a significant risk factor for developing the disease.

RA is often found together with other autoimmune diseases, such as thyroiditis, vitiligo, diabetes mellitus (DM), and Sjogren syndrome. The incidence of DM is increased in patients with RA and a meta-analysis found a pooled relative risk of 1.23 (Tian et al., 2021). In our study group, 38 RA patients had DM (OR = 1.20, p = 0.52), indicating that the presence of DM was not a significant factor associated with the diagnosis of RA. Vitiligo is a common autoimmune skin disease, seen in 0.5-2% of the population, frequently seen together with many other diseases such as thyroiditis, DM, pernicious anemia, psoriasis and RA. In addition to RA, it is also frequently found together with other rheumatological diseases (Dahir & Thomsen, 2018). This study showed that although 7 out of 8 patients diagnosed with vitiligo had RA, the presence of vitiligo was not a statistically significant factor in the diagnosis of RA (OR = 4.43, p = 0.16). Autoimmune thyroiditis (AT) is the most common human autoimmune disorder, with a prevalence of around 5%. AT patients often have another autoimmune disorder, such as RA, pernicious anemia, systemic lupus erythematosus, vitiligo. The prevalence of AT in RA patients is increased by 1–6-fold, and the prevalence of RA in AT patients is increased by 1–3-fold (Conigliaro et al., 2020). In this study group, 47 RA patients had AT (OR = 3.73, p < 0.001), indicating that a history of AT or the presence of thyroid autoantibodies was a significant factor associated with the diagnosis of RA. The incidence of HT is increased in patients with RA and negatively affects the course of the disease, and one study found the incidence of HT in patients with RA to be 59.5% (Hadwen et al., 2021; Morović-Vergles et al., 2013). Hypertension was present in 81 of our RA patients (OR = 2.06, p = 0.003), which may be related to the age group in which both diseases are commonly observed (Hadwen et al., 2021; Morović-Vergles et al., 2013).

RFs are positive in 60-80% of RA patients, and may also be positive in other autoimmune diseases, during the course of infections and in old age. RF specificity increases at higher titers (RF>50 IU/ml). anti-CCPs are positive in 70-90% of RA patients, with higher specificity (90-95%), and are rarely positive in other diseases or in healthy individuals (anti-CCP may rarely be positive in diseases such as systemic lupus erythematosus, Sjogren's disease, psoriatic arthritis and tuberculosis (Conigliaro et al., 2016; Lima et al., 2013)). In general, anti-CCP has better diagnostic value than RF in terms of sensitivity and specificity (Ingegnoli et al., 2013; Nielsen et al., 2012). The RF cut off level for our hospital laboratory is 15 IU/mL and the anti-CCP cut off level is 5 U/mL. Using this cut-off value, our study showed that RF positivity is a significant risk factor in the diagnosis of RA (OR = 12.3, p < 0.001), similarly, anti-CCP positivity was found to be a stronger

diagnostic marker for RA than RF (OR = 38.4, $p < 0.001$).

In recent years, successful studies have been conducted to diagnose RA using transfer learning and CNN, a deep learning method, from radiographs of patients' hands (Ma et al., 2024; Üreten & Maraş, 2022). Similarly, machine learning methods have been conducted to diagnose RA using patient demographic characteristics, clinical features and laboratory results. For disease diagnosis, Sundaramurthy et al. conducted a study on 250 healthy individuals and 750 RA patients, achieving 94% accuracy, 90.80% PPV, and 94.70% ROC. Their study did not rank the importance of features in disease diagnosis. Matsuo H and colleagues, in their study on predicting RA relapse over 2 years of follow-up, examined 210 patients and used 73 ultrasound features and 54 blood test results. Using the XGBoost classifier, they obtained an AUC score of 0.747. The study did not rank the features effective in the occurrence of relapses (Matsuo et al., 2022; Sundaramurthy et al., 2020).

Feature importance indicates how much each feature contributes to the prediction of the model, and determines the degree of usefulness of a given feature for an existing model and prediction. Feature importance helps in better understanding data, making analyses easier by revealing the factors with the greatest impact on results (Castiglioni et al., 2021; Greener et al., 2022). The first 7 features in the feature importance ranking were anti-CCP, RF, age, Hb, CRP, neutrophil and lymphocytes. In the study conducted with dataset-2, from which serological test results, RF and anti-CCP were excluded, the important features in the diagnosis of RA were listed as follows: morning_stiffness, ht, thyroiditis, family_ra, crp, age and gender. Models developed with dataset-2 can be applied in every center, but show lower performance in all statistical metrics. For example, the random forest model developed with dataset-1 yielded 88% accuracy and 92% AUC, while the study conducted with dataset-2 yielded 74% accuracy and 73% AUC. This model needs to be improved with a larger dataset.

This study had some limitations. First, the number of patients was insufficient, some patients had been previously diagnosed RA and were taking medication for RA. For this reason, the levels of acute phase reactants such as CRP and ESR may be within normal limits. In RA, elevated CRP levels due to inflammation are an important finding and are used in diagnosis. In the ranking of feature importance, CRP was ranked 5th. This may have reduced the performance of these models. Second, this study was performed on patients who were evaluated and examined in the rheumatology outpatient clinic and the number of patients with RA was higher than in the non-RA group (260/161). This may be the reason why the RF and anti-CCP Odds ratios used in the diagnosis of the disease are high. Due to the high workload and limited number of rheumatology outpatient clinics, it may be difficult for RA patients complaining of pain in hand joints to access these centers. Another limitation of this study was that there were no patients examined by a general practitioner, orthopedist, or physical therapy specialist with complaints of pain in the hands. Model performance can be improved and bias prevented through multicenter studies. They also ensure the generalizability of the developed models.

IV. CONCLUSION

In recent years, machine learning and ensemble learning models have been successfully used in many studies in the medical field. In ensemble learning models, several machine learning algorithms are used together. In this study on RA diagnosis and prediction, all three ensemble learning models showed similar performance and outperformed the machine learning algorithms. This proposed model can be improved with new multicenter studies including more patients, encompassing rheumatology, orthopedics, and physical therapy centers, thus ensuring the generalizability of the developed models. RF and anti-CCP serological tests used in RA diagnosis cannot be performed in every center. The development of the model with dataset-2 will prevent unnecessary effort for patients to undergo these tests. By developing the proposed models with more patients, unnecessary rheumatology referrals are avoided, and the developed model proves beneficial even in rural areas where serological tests such as RF and anti-CCP are unavailable. Hand osteoarthritis, characterized by pain and limited range of motion in the hand joints and increasing in frequency with age, is treated with physical therapy and nonsteroidal anti-rheumatic drugs. Treatment of RA involves immunosuppressive drugs such as Methotrexate and Sulfasalazin, and biological therapies such as anti-tumor necrotizing drugs. Machine learning models can be developed to differentiate between these two diseases, which have different treatments and courses.

DECLARATIONS

Acknowledgements: The authors do not wish to acknowledge any individual or institution.

Author Contributions: Data collection Y.M., E.A. and K.O., data analysis K.Ü., and K.O., writing—original draft, K.Ü., and Y.M., writing—review and editing, K.Ü., and E.A. All authors have read and approved the final version of it.

Conflict of Interest: The authors declare that they have no conflicts of interest.

Supporting Institutions: This research received no external funding.

Ethical Approval: An ethical approval certificate was obtained from the local ethic commitee, Ankara Bilkent City Hospital, Clinical Research Ethics Committee No. 1, date : 16.08.2023, no : E1-23-3650. This study is a retrospective study and ethics committee approval was obtained.

Plagiarism Statement: This article has been evaluated for plagiarism and no instances of plagiarism were detected.

Use of AI Tools: The authors declare that no Artificial Intelligence (AI) tools were used in the creation of this article.

REFERENCES

Abdelhafiz, D., Baker, T., Glascow, D., & Abdelhafiz, A. (2023). Biomarkers for the diagnosis and treatment of rheumatoid arthritis – a systematic review. *Postgraduate Medicine*, *135*(3), 214–223. https://doi.org/10.1080/00325481.2022.2052626

Ali, M. M., Paul, B. K., Ahmed, K., Bui, F. M., Quinn, J. M. W., & Moni, M. A. (2021). Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison. *Computers in Biology and Medicine*, *136*, Article 104672. https://doi.org/10.1016/j.compbiomed.2021.104672

Almutairi, K., Nossent, J., Preen, D., Keen, H., & Inderjeeth, C. (2021). The global prevalence of rheumatoid arthritis: A meta-analysis based on a systematic review. *Rheumatology International*, *41*(5), 863–877. ttps://doi.org/10.1007/s00296-020-04731-0

Azmi, S. S., & Baliga, S. (2020). An overview of boosting decision tree algorithms utilizing AdaBoost and XGBoost boosting strategies. *International Research Journal of Engineering and Technology*, *7*(5), 6867–6870.

Azur, M. J., Stuart, E. A., Frangakis, C., & Leaf, P. J. (2011). Multiple imputation by chained equations: What is it and how does it work? *International Journal of Methods in Psychiatric Research*, *20*(1), 40–49. https://doi.org/10.1002/MPR.329

Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. *Electronics*, *8*(8), Article 832. https://doi.org/10.3390/electronics8080832

Castiglioni, I., Rundo, L., Codari, M., Di Leo, G., Salvatore, C., Interlenghi, M., Gallivanone, F., Cozzi, A., D'Amico, N. C., & Sardanelli, F. (2021). AI applications to medical images: From machine learning to deep learning. *Physica Medica*, *83*, 9–24. https://doi.org/10.1016/j.ejmp.2021.02.006

Conigliaro, P., Chimenti, M. S., Triggianese, P., Sunzini, F., Novelli, L., Perricone, C., & Perricone, R. (2016). Autoantibodies in inflammatory arthritis. *Autoimmunity Reviews*, *15*(7), 673–683. https://doi.org/10.1016/j.autrev.2016.03.003

Conigliaro, P., D'Antonio, A., Pinto, S., Chimenti, M. S., Triggianese, P., Rotondi, M., & Perricone, R. (2020). Autoimmune thyroid disorders and rheumatoid arthritis: A bidirectional interplay. *Autoimmunity Reviews*, *19*(6), Article 102529. https://doi.org/10.1016/j.autrev.2020.102529

Dahir, A. M., & Thomsen, S. F. (2018). Comorbidities in vitiligo: Comprehensive review. *International Journal of Dermatology*, *57*(10), 1157–1164. https://doi.org/10.1111/ijd.14055

Deane, K. D., & Holers, V. M. (2019). The natural history of rheumatoid arthritis. *Clinical Therapeutics*, *41*(7), 1256–1269. https://doi.org/10.1016/j.clinthera.2019.04.028

Dong, X., Yu, Z., Cao, W., Shi, Y., & Ma, Q. (2020). A survey on ensemble learning. *Frontiers of Computer Science*, *14*(2), 241–258. https://doi.org/10.1007/s11704-019-8208-z

Dritsas, E., & Trigka, M. (2022). Machine learning techniques for chronic kidney disease risk prediction. *Big Data and Cognitive Computing*, *6*(3), Article 98. https://doi.org/10.3390/bdcc6030098

Emre, İ. E., Erol, N., Ayhan, Y. İ., Özkan, Y., & Erol, Ç. (2019). The analysis of the effects of acute rheumatic fever in childhood on cardiac disease with data mining. *International Journal of Medical Informatics*, *123*, 68–75. https://doi.org/10.1016/j.ijmedinf.2018.12.009

Fallahi, P., Ferrari, S. M., Ruffilli, I., Elia, G., Biricotti, M., Vita, R., Benvenga, S., & Antonelli, A. (2016). The association of other autoimmune diseases in patients with autoimmune thyroiditis: Review of the literature and report of a large series of patients. *Autoimmunity Reviews*, *15*(12), 1125–1128. https://doi.org/10.1016/j.autrev.2016.09.009

Faruque, M. F., Asaduzzaman, & Sarker, I. H. (2019). Performance analysis of machine learning techniques to predict diabetes mellitus. In *2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)* (pp. 1–4). https://doi.org/10.1109/ECACE.2019.8679365

Frisell, T., Saevarsdottir, S., & Askling, J. (2016). Family history of rheumatoid arthritis: An old concept with new developments. *Nature Reviews Rheumatology*, *12*(6), 335–343. https://doi.org/10.1038/nrrheum.2016.52

Greener, J. G., Kandathil, S. M., Moffat, L., & Jones, D. T. (2022). A guide to machine learning for biologists. *Nature Reviews Molecular Cell Biology*, *23*(1), 40–55. https://doi.org/10.1038/s41580-021-00407-0

Hadwen, B., Stranges, S., & Barra, L. (2021). Risk factors for hypertension in rheumatoid arthritis patients–A systematic review. *Autoimmunity Reviews*, *20*(4), Article 102786. https://doi.org/10.1016/j.autrev.2021.102786

Ingegnoli, F., Castelli, R., & Gualtierotti, R. (2013). Rheumatoid factors: Clinical applications. *Disease Markers*, *35*(6), 727–734. https://doi.org/10.1155/2013/726598

Janiesch, C., Zschech, P., & Heinrich, K. (2021). Machine learning and deep learning. *Electronic Markets*, *31*(3), 685–695. https://doi.org/10.1007/S12525-021-00475-2

Jindal, G., Bansal, S., Gupta, N., Singh, S. K., Gahukar, S., & Kumar, A. (2021). Comparison of ultrasonography and X-Rays for the diagnosis of synovitis and bony erosions in small joints of hands in early rheumatoid arthritis: A prospective study. *Mædica - a Journal of Clinical Medicine*, *16*(1), 22–28. https://doi.org/10.26574/maedica.2020.16.1.22

Lima, I., Oliveira, R. C., Atta, A., Marchi, S., Barbosa, L., Reis, E., Reis, M. G., & Santiago, M. B. (2013). Antibodies to citrullinated peptides in tuberculosis. *Clinical Rheumatology*, *32*(5), 685–687. https://doi.org/10.1007/s10067-013-2173-y

Luo, Y. (2022). Evaluating the state of the art in missing data imputation for clinical data. *Briefings in Bioinformatics*, *23*(1), 1–9. https://doi.org/10.1093/bib/bbab489

Ma, Y., Pan, I., Kim, S. Y., Wieschhoff, G. G., Andriole, K. P., & Mandell, J. C. (2024). Deep learning discrimination of rheumatoid arthritis from osteoarthritis on hand radiography. *Skeletal Radiology*, *53*(2), 377–383. https://doi.org/10.1007/s00256-023-04408-2

Mahesh, B. (2018). Machine learning algorithms-a review. *International Journal of Science and Research, 9*(1), 381–386. https://doi.org/10.21275/ART20203995

Matsuo, H., Kamada, M., Imamura, A., Shimizu, M., Inagaki, M., Tsuji, Y., Hashimoto, M., Tanaka, M., Ito, H., & Fujii, Y. (2022). Machine learning-based prediction of relapse in rheumatoid arthritis patients using data on ultrasound examination and blood test. *Scientific Reports*, *12*, Article 7224. https://doi.org/10.1038/s41598-022-11361-y

Mendoza-Pinto, C., Sánchez-Tecuatl, M., Berra-Romani, R., Maya-Castro, I. D., Etchegaray-Morales, I., Munguía-Realpozo, P., Cárdenas-García, M., Arellano-Avendaño, F. J., & García-Carrasco, M. (2024). Machine learning in the prediction of treatment response in rheumatoid arthritis: A systematic review. *Seminars in Arthritis and Rheumatism*, *68*, Article 152501. https://doi.org/10.1016/j.semarthrit.2024.152501

Morović-Vergles, J., Šalamon, L., Marasović-Krstulović, D., Kehler, T., Šakić, D., Badovinac, O., Vlak, T., Novak, S., Štiglić-Rogoznica, N., Hanih, M., Bedeković, D., Grazio, S., Kadojić, M., Milas-Ahić, J., Prus, V., Stamenković, D., Šošo, D., Anić, B., Babić-Naglić, D., & Gamulin, S. (2013). Is the prevalence of arterial hypertension in rheumatoid arthritis and osteoarthritis associated with disease? *Rheumatology International*, *33*(5), 1185–1192. https://doi.org/10.1007/s00296-012-2522-1

Muhammad, L. J., Algehyne, E. A., Usman, S. S., Ahmad, A., Chakraborty, C., & Mohammed, I. A. (2021). Supervised machine learning models for prediction of COVID-19 infection using epidemiology dataset. *SN Computer Science*, *2*(1), Article 11. https://doi.org/10.1007/s42979-020-00394-7

Mun, S., Lee, J., Park, M., Shin, J., Lim, M.-K., & Kang, H.-G. (2021). Serum biomarker panel for the diagnosis of rheumatoid arthritis. *Arthritis Research & Therapy*, *23*(1), Article 31. https://doi.org/10.1186/s13075-020-02405-7

Mutasa, S., Sun, S., & Ha, R. (2020). Understanding artificial intelligence based radiology studies: What is overfitting? *Clinical Imaging*, *65*, 96–99. https://doi.org/10.1016/j.clinimag.2020.04.025

Nielsen, S. F., Bojesen, S. E., Schnohr, P., & Nordestgaard, B. G. (2012). Elevated rheumatoid factor and long term risk of rheumatoid arthritis: A prospective cohort study. *BMJ*, *345*(7878), Article e5244. https://doi.org/10.1136/bmj.e5244

Otón, T., & Carmona, L. (2019). The epidemiology of established rheumatoid arthritis. *Best Practice & Research Clinical Rheumatology*, *33*(5), Article 101477. https://doi.org/10.1016/j.berh.2019.101477

Radu, A.-F., & Bungau, S. G. (2021). Management of rheumatoid arthritis: An overview. *Cells*, *10*(11), Article 2857. https://doi.org/10.3390/cells10112857

Romão, V. C., & Fonseca, J. E. (2021). Etiology and risk factors for rheumatoid arthritis: A state-of-the-art review. *Frontiers in Medicine*, *8*, Article 689698. https://doi.org/10.3389/FMED.2021.689698

Sagi, O., & Rokach, L. (2018). Ensemble learning: A survey. *WIREs Data Mining and Knowledge Discovery*, *8*(4), Article e1249. https://doi.org/10.1002/widm.1249

Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science*, *2*(3), Article 160. https://doi.org/10.1007/S42979-021-00592-X

Sundaramurthy, S., Saravanabhavan, C., & Kshirsagar, P. (2020). 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). https://doi.org/10.1109/DASA51403.2020.9317253

Tian, Z., McLaughlin, J., Verma, A., Chinoy, H., & Heald, A. H. (2021). The relationship between rheumatoid arthritis and diabetes mellitus: A systematic review and meta-analysis. *Cardiovascular Endocrinology and Metabolism*, *10*(2), 125–131. https://doi.org/10.1097/XCE.0000000000000244

Üreten, K., & Maraş, H. H. (2022). Automated classification of rheumatoid arthritis, osteoarthritis, and normal hand radiographs with deep learning methods. *Journal of Digital Imaging*, *35*(2), 193–199. https://doi.org/10.1007/s10278-021-00564-w

Wu, C. T., Lo, C. L., Tung, C. H., & Cheng, H. L. (2020). Applying data mining techniques for predicting prognosis in patients with rheumatoid arthritis. *Healthcare*, *8*(2), Article 85. https://doi.org/10.3390/healthcare8020085