

Early Risk Prediction of Rheumatoid Arthritis Using a Hybrid Restricted Boltzmann Machine

Riska Rahma Yanti¹, Siti Amiroch², Awawin Mustana Rohmah³, Dicka Yale Kardono⁴

Abstract—Rheumatoid arthritis (RA) is a chronic autoimmune disease that requires early detection to prevent permanent joint damage. The main challenge in predicting RA risk lies in the complexity of clinical data, which is multidimensional and often imbalanced. This study employs a Hybrid Restricted Boltzmann Machine (H-RBM) approach to improve the accuracy of early RA risk prediction. The RBM method is used as a feature extractor to map non-linear relationships within clinical data, which is then combined with three classification algorithms: Support Vector Machine (SVM), Random Forest (RF), and Logistic Regression (LR). The research data was sourced from Kaggle, comprising 4,425 samples that were balanced using the Synthetic Minority Oversampling Technique (SMOTE). Testing was conducted with three data split ratios (70:30, 80:20, and 90:10) using GridSearchCV for hyperparameter tuning. The experimental results show that the RBM-SVM model dominates absolutely with the highest performance at the 90:10 ratio, yielding an accuracy of 99.82%, an AUC score of 1.0000, and a recall of 100%. The perfect recall rate indicates that the model is capable of identifying all positive patients without any missed cases (zero false negatives). Meanwhile, the RBM-RF and RBM-LR models achieved optimal performance at an 80:20 ratio with accuracies of 93.95% and 83.42%, respectively. Overall, RBM integration proved significant in extracting more informative latent information compared to raw clinical features, thereby making a crucial contribution to a precise RA early detection decision support system.

Keywords : rheumatoid arthritis, prediction, restricted Boltzmann machine, deep learning

I. INTRODUCTION

RHEUMATOID arthritis (RA) is a chronic autoimmune disease that affects the joints and can cause permanent damage if not detected and treated early [1]. Rheumatoid arthritis occurs when the immune system, which normally helps protect the body from infection and disease, begins to attack its own tissues. This disease causes pain, swelling, stiffness, and loss of function in the joints [2]. Despite its serious impact, awareness of RA remains relatively low compared to other chronic diseases, leading to frequent neglect of early detection. According to 2019 data from the World Health Organization (WHO), nearly 18 million people worldwide suffer from RA, with approximately 70% of rheumatoid arthritis patients being women, and 55% of them over the age of 55; by 2022, more than 25% experienced disability due to rheumatoid arthritis [3]. Many members of the general public are unaware of the early symptoms of RA; nonspecific

symptoms such as fatigue, stiffness, and mild joint pain in the morning are often mistaken for normal occurrences, thereby delaying proper diagnosis and treatment [4].

Machine learning (ML) offers a potential solution for RA risk prediction based on clinical data [5]. Although conventional models can achieve high prediction accuracy, they still have limitations in capturing nonlinear and complex relationships within multidimensional data. Therefore, deep learning approaches are considered more effective because they can automatically learn data representations through hidden layers [6]. Deep learning is a branch of artificial intelligence capable of automatically learning data representations through multiple hidden layers [7]. Commonly used deep learning methods include Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Deep Belief Networks (DBNs), Multi-Layer Perceptrons (MLPs), and Restricted Boltzmann Machines (RBMs). However, CNNs and RNNs are generally better suited for image and sequential data, while DBNs and MLPs have relatively more complex architectures. In this context, RBMs offer a simpler yet effective alternative, as they can extract features from high-dimensional data by modeling the relationships between visible and hidden units.

The Restricted Boltzmann Machine (RBM) is an energy-based probabilistic model designed to extract features from high-dimensional data by modeling the relationships between visible and hidden units. RBM's capability in feature learning makes it one of the suitable methods for supporting early RA risk prediction systems [8]. In a previous study using the Sustainable Physical Healthcare Pattern Recognition (SPHR) approach with a hybrid feature model based on RBM, the best results were achieved with an accuracy of 93.33% on the mHealth dataset and 92.50% on HuGaDB [9]. Research using a deep learning approach based on RBM for breast cancer classification yielded higher accuracy compared to conventional machine learning methods [10]. Based on previous studies, it is evident that RBM in deep learning is effective in the classification process and is highly recommended for use as a decision-support method for disease-related data, such as image, sensory, and clinical data. To provide an overview of the latest research developments and to highlight the significance and novelty of the proposed study, Table I presents a comparison of several previous studies that have utilized RBM in various fields, including the methods used, research subjects, results obtained, limitations, and the contributions of this study.

*This work was not supported by any organization

^{1,2,3,4} Faculty of Science and Technology, Universitas Islam Darul 'Ulum Lamongan, Indonesia

siti.amiroch@unisda.ac.id

Manuscript received April 29, 2026; accepted July 17, 2026.

TABLE I
STATE-OF-THE-ART COMPARISON

Authors	Method	Application Domain	Performance	Research Gap / Novelty
Yudicy Amelia et al. (2018)	Machine Learning and RBM-based Deep Learning	Breast Cancer Dataset	Deep Learning Accuracy: 98.5% and Machine Learning Accuracy: 97%	The study focused on breast cancer classification. This research applies RBM to autoimmune disease prediction using RA clinical data.
Mahida Javeed et al. (2021)	Hybrid RBM with DBN, AdaBoost, and RF	mHealth and HuGaDB	DBN mHealth dataset Accuracy: 93.33% and DBN HuGaDB dataset Accuracy: 92.50%	The work utilized sensor-based datasets. In contrast, this study employs clinical data for early rheumatoid arthritis risk prediction.
Haviluddin et al. (2025)	Restricted Boltzmann Machine (RBM)	Respiratory Disease Dataset	RBM Accuracy: 98%	RBM was implemented as a standalone classifier for respiratory diseases. This study integrates RBM with SVM, RF, and LR to compare classification performance on RA clinical data.
This Study	Hybrid RBM-SVM, RBM-RF, and RBM-LR	Rheumatoid Arthritis Dataset	RBM-SVM Accuracy: 99.82%, Recall: 100%, AUC: 1.0000	Integrates RBM as a latent feature extractor with multiple classifiers and evaluates different train-test split ratios for early rheumatoid arthritis risk prediction.

Based on Table I, previous studies have shown that RBM is effective as a feature extraction method in various domains, such as health sensor data and cancer classification. However, the use of RBM as a feature extractor for clinical data on rheumatoid arthritis in combination with various classification algorithms remains limited. Furthermore, previous studies have generally focused on a single classification model, thus failing to provide a comprehensive evaluation of the most suitable algorithms for utilizing the latent features extracted by RBM. Therefore, the novelty of this study lies in integrating RBM with three classification algorithms, namely Support Vector Machine (SVM), Random Forest (RF), and Logistic Regression (LR), as well as evaluating their performance under different train-test split ratios to identify the most effective model for early-stage Rheumatoid Arthritis risk prediction. This approach combines RBM's feature extraction capabilities with the strengths of each classifier: SVM effectively handles nonlinear patterns and high-dimensional data, RF manages data complexity while reducing overfitting, and LR provides a simple and interpretable model. Consequently, the proposed framework is expected to improve the accuracy, sensitivity, and specificity of early RA detection.

II. METHODOLOGY

The dataset used in this study was sourced from the Rheumatic and Autoimmune Disease Dataset, which is publicly available on the Kaggle platform. The original dataset consists of 12,085 patient records covering various types of autoimmune diseases. Since this study focuses on the classification of rheumatoid arthritis (RA), a filtering process was performed using Microsoft Excel to select data from patients diagnosed with RA as well as healthy controls. The filtering process yielded 4,452 records, which were subsequently used in the model training and testing processes. The dataset used consists of 14 input features and 1 target feature. The input features represent patient characteristics, while the target feature indicates the diagnosis result used as the classification label. The target variable is binary—Rheumatoid Arthritis (RA) or Healthy—making this a binary classification problem. The input features used include: age, gender,

Erythrocyte Sedimentation Rate (ESR), C-Reactive Protein (CRP), Rheumatoid Factor (RF), Anti-Cyclic Citrullinated Peptide (Anti-CCP), HLA-B27, Antinuclear Antibody (ANA), Anti-Sjögren's Syndrome-Related Antigen (Anti-Ro), Anti-Sjögren's Syndrome-Related Antigen (Anti-La), Anti-Double-Stranded DNA (Anti-dsDNA), Anti-Smith Antibody (Anti-sm), C3, and C4, while the target feature is Diagnosis, consisting of two classes: Rheumatoid Arthritis (RA) and Healthy. The use of this filtered dataset allows the classification model to learn characteristics that differentiate RA patients from healthy individuals without being influenced by data from other autoimmune diseases.

A. Qualitative Analysis Methods

The stages of this research are illustrated in a flowchart shown in Figure.1 with the following explanation:

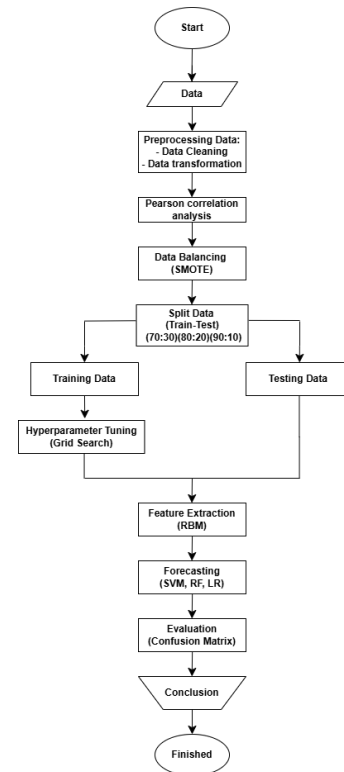


Fig. 1. Research Flowchart

B. Data Preprocessing

Preprocessing is a component of data preparation that encompasses all types of processing applied to raw data to prepare it for subsequent data processing procedures [12]. This process involves several stages, including data cleaning, data balancing, and normalization.

C. Pearson Correlation Analysis

Pearson correlation analysis is used to determine the strength and direction of the linear relationship between two variables on an interval or ratio scale, with the aim of identifying the extent of the relationship between the independent

variable and the target variable [11]. The Pearson correlation coefficient is formulated as follows:

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2}} \quad (1)$$

where:

- r = Pearson correlation coefficient
- X_i = the i -th value of the independent variable
- Y_i = the i -th value of the dependent variable
- $\bar{X}$ = mean of variable X
- $\bar{Y}$ = mean of variable Y

The interpretation of correlation values is presented in Table II as follows:

TABLE II
CORRELATION VALUES

Value	Interpretation
0.00 – 0.19	Very weak relationship
0.20 – 0.39	Weak relationship
0.40 – 0.59	Moderate relationship
0.60 – 0.79	Strong relationship
0.80 – 1.00	Very strong relationship

Based on Table II, features with correlation values that are very low (close to zero) with respect to the target variable indicate a very weak relationship and may be considered for removal to improve model efficiency.

D. Data Balancing

Data balancing is a technique used to address imbalances in datasets (imbalanced datasets), which is a condition where the number of samples in one class is significantly greater than in the others [14]. In this study, the Synthetic Minority Oversampling Technique (SMOTE) method was used to address data imbalance. SMOTE is a commonly used method for balancing the distribution by generating synthetic data in the minority class. This method was chosen based on its ability to improve data balance and produce more accurate model performance. The formula for determining synthetic data is as follows:

$$X_{syn} = X_i + (X_{knn} - X_i) \times \sigma \quad (2)$$

where:

- X_{syn} = synthetic data
- X_i = original data point
- X_{knn} = nearest neighbor data
- σ = random value between 0 and 1

E. Data Splitting

Data splitting is the process of dividing a dataset into several parts for the purposes of training and testing machine learning models [15]. In this study, the data was divided into two subsets: training data and testing data, using commonly employed split ratios of 70:30, 80:20, and 90:10 [16]. The larger portion of the data is used to train the model, while the remainder is used to evaluate the model's performance. Subsequently, the best results from each of these split ratios are selected as the basis for determining the RBM prediction model combined with a classification algorithm.

F. Hyperparameter Tuning

Hyperparameter tuning is the process of optimizing parameters that the model does not learn during training, such as the learning rate, the number of hidden layers, or the maximum depth in a decision tree. Hyperparameter tuning aims to optimize performance metrics such as accuracy, F1-score, or the loss function by experimenting with different combinations of hyperparameters [17].

G. Early Risk Prediction Mechanism

Early risk prediction in this study refers to the identification of patients showing early signs of rheumatoid arthritis (RA) based on clinical data obtained during the initial examination. This study does not predict future disease onset but rather classifies risk using patients' clinical data. The input data used in the prediction process are described in the Dataset subsection as follows.

TABLE III
DATA INPUT

Feature	Data Type	Sample Value
Age	Numerical	39
Gender	Categorical	Female/Male
ESR	Numerical	26
...	...	...
C3	Numerical	100
C4	Numerical	66

Before classification, the data first undergoes a preprocessing stage, which includes handling missing values, encoding, feature standardization, and data balancing using the SMOTE method. Next, RBM is used to extract latent features, which then serve as input for the SVM, Random Forest (RF), and Logistic Regression (LR) algorithms. The output of the proposed system is a binary prediction indicating whether a patient belongs to the Normal (0) or Rheumatoid Arthritis Risk (1) class. The format of the prediction output table is presented in Table IV.

TABLE IV
SAMPLE OUTPUT OF EARLY RA RISK PREDICTION

Probability Score	Actual Disease	Prediction
0.032	Normal	Normal
0.087	Normal	Normal
0.996	RA	RA
0.981	RA	RA
0.945	RA	RA

Table IV shows an example of the output from the early RA risk prediction model. Probability values close to 0 indicate the Normal class, while values close to 1 indicate the RA class. The agreement between the actual labels and the prediction results demonstrates the model's ability to accurately classify the two classes.

H. Restricted Boltzmann Machine (RBM)

The Restricted Boltzmann Machine (RBM) is a probabilistic model in deep learning used to learn data representations. This

model is capable of extracting features from input data through an unsupervised learning approach [18]. In this study, feature extraction will be performed, and the resulting latent features will be used for prediction using a classification algorithm. The RBM uses an energy score to determine how well the configuration between the input units (v) and the hidden feature units (h) matches [19].

$$E(v, h, \theta) = - \sum_i v_i a_i - \sum_j h_j b_j - \sum_{i,j} v_i W_{ij} h_j \quad (3)$$

where:

- $E(v, h, \theta)$ = energy value of the unit configuration (v) and (h),
- v_i = visible unit
- h_j = hidden unit
- a_i, b_j = bias
- W_{ij} = the weight connecting the visible unit to the hidden unit.

Based on this function, the activation probability of a hidden unit can be expressed as

$$p(h_k = 1 | v) = \sigma \left(a_k + \sum_j W_{kj} v_j \right), \quad (4)$$

where:

- $p(h_k = 1 | v)$ = probability that the k -th hidden unit is activated (takes the value 1) given the visible input vector v ;
- a_k = bias of the k -th hidden unit;
- W_{kj} = weight connecting the j -th visible unit to the k -th hidden unit;
- v_j = value of the j -th visible unit; and
- $\sigma(\cdot)$ = sigmoid activation function.

The sigmoid activation function maps a real-valued input into the interval $[0, 1]$, allowing the output to be interpreted as a probability. It is defined as

$$\sigma(x) = \frac{1}{1 + e^{-x}}, \quad (5)$$

where:

- $\sigma(x)$ = output of the sigmoid activation function;
- x = input to the activation function; and
- e = Euler's number, approximately equal to 2.71828.

According to Equation 5, as the input value x increases, the output $\sigma(x)$ approaches 1. Conversely, as x decreases, the output $\sigma(x)$ approaches 0. Therefore, the sigmoid function transforms real-valued inputs into values within the interval $[0, 1]$, which can be interpreted as probabilities.

The conditional independence property of RBMs allows each hidden unit to be computed efficiently, thereby simplifying the process of learning high-dimensional data representations. The RBM training process is performed by maximizing the log-likelihood through parameter gradient calculation, which is formulated as:

$$\frac{\partial \log p(v)}{\partial W_{ij}} = \langle v_i h_j \rangle_{data} - \langle v_i h_j \rangle_{model} \quad (6)$$

Since the expectation of the model distribution is difficult to calculate exactly, the Contrastive Divergence (CD-k) algorithm is used with a weight update rule given by the formula:

$$\Delta W = \eta (\langle v h \rangle_{data} - \langle v h \rangle_{recon}) \quad (7)$$

where:

- ΔW = Weight change, i.e., how much the weights are updated in a single iteration
- η = learning rate, the model's learning rate
- $\langle v h \rangle_{data}$ = expected product of the hidden units when the input is real data (data distribution).
- $\langle v h \rangle_{recon}$ = expected product of $v_i h_j$ ' when the input is reconstructed from the hidden layer.

After the training process, the RBM generates latent features using the formula:

$$H = P(h = 1 | v) \quad (8)$$

which is subsequently used as input for classification models. Thus, the RBM acts as a feature extractor capable of improving prediction performance on high-dimensional and complex data.

I. Classification Algorithms

Classification is a technique in machine learning that aims to categorize or classify data into predetermined labels or classes. Classification algorithms work by learning patterns from labeled data during the training phase, then using those patterns to classify new, unlabeled data[20]. This study will compare three classification algorithms: Support Vector Machine (SVM), Random Forest (RF), and Logistic Regression (LR).

1) *Support Vector Machine (SVM)*: SVM is a supervised learning method used for classification and regression [18]. The decision function is:

$$f(H) = w^T H + b \quad (9)$$

where:

- w^T = transpose of the weight vector w
- b = bias
- H = hidden pattern vector from the RBM

Prediction:

$$\hat{y} = \begin{cases} 1, & w^T H + b \geq 0 \\ 0, & w^T H + b < 0 \end{cases} \quad (10)$$

The SVM model is trained by minimizing the objective function using equation (10) as follows:

$$\min_{w,b} \left(\frac{1}{2} \|w\|^2 + C \sum_{i=1}^n \xi_i \right) \quad (11)$$

where:

- $\|w\|^2$ = weighted least squares (margin maximization)
- ξ_i = a slack variable to account for classification error
- C = a regularization parameter that controls the trade-off between margin and classification error
- n = the number of training data points

2) *Random Forest (RF)*: Random Forest is an algorithm that uses the bagging technique with decision trees as its base model. This method works by randomly building a number of decision trees from different subsets of the data, then combining the predictions from each tree to produce a final decision [22]. Model Input:

$$H = f_{\text{RBM}}(X) \quad (12)$$

For each t th decision tree, a prediction is obtained:

$$T_1(H), T_2(H), \dots, T_T(H) \quad (13)$$

where:

- $T_t(\cdot)$ = the t -th decision tree
- T = the number of trees in the Random Forest

The final prediction is determined based on a majority vote:

$$\hat{y} = \text{majority vote}(T_1(H), \dots, T_T(H)) \quad (14)$$

3) *Logistic Regression (LR)*: Logistic regression (LR) is a statistical method that falls under supervised machine learning algorithms and is used to model the relationship between one or more independent variables and a categorical dependent variable (typically binary: 0 or 1, yes/no) [23].

$$P(y = 1|H) = \frac{1}{1 + e^{-(w^T H + b)}} \quad (15)$$

where:

- $P(y = 1 | H)$ = data probability
- H = hidden pattern vector extracted by RBM
- w = LR model weight vector
- b = bias
- $w^T H$ = dot product of the weights and the hidden pattern
- e = exponential number

The classification decision is determined based on a threshold value of 0.5, as shown in Equation (16) below:

$$\hat{y} = \begin{cases} 1, & P(y = 1 | H) \geq 0.5 \\ 0, & \text{otherwise} \end{cases} \quad (16)$$

where:

- $\hat{y}$ = predicted class
- 1 = abnormal class
- 0 = normal class

The loss function minimized in Logistic Regression is the Binary Cross-Entropy Loss, which is formulated as in equation (16):

$$L = - \sum_{i=1}^n [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)] \quad (17)$$

where:

- L = the value of the loss function
- n = the number of training data points
- y_i = the actual label of the i th data point
- p_i = the predicted probability $P(y = 1 | H)$

J. Model Evaluation

Model performance is evaluated using a *Confusion Matrix* to determine the number of prediction errors, such as False Positives and False Negatives [20]. In binary classification, this matrix consists of four main components, namely:

- 1) **True Positive (TP)**: a condition where both the predicted result and the actual value indicate the positive class
- 2) **True Negative (TN)**: the condition where both the predicted result and the actual value indicate the negative class
- 3) **False Positive (FP)**: the condition where the model predicts the positive class, but the actual value is negative (*Type I Error*)
- 4) **False Negative (FN)**: a condition where the model predicts a negative class, but the actual value is positive (*Type II Error*)

The TP, TN, FP, and FN values can be used to calculate the Accuracy, Precision, Recall, and F1-Score for each class. Metrics such as accuracy, precision, recall, and F1-score provide information about the model's performance. These metrics can be calculated using the following equations:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \quad (18)$$

$$\text{Precision} = \frac{TP}{TP + FP} \times 100\% \quad (19)$$

$$\text{Recall} = \frac{TP}{TP + FN} \times 100\% \quad (20)$$

$$F1 - \text{Score} = \frac{2 \times (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}} \times 100\% \quad (21)$$

The combination of these five evaluation metrics is crucial for ensuring clinical safety in the early risk prediction of rheumatoid arthritis (RA). Through this integration, the model is not only evaluated based on overall accuracy but is also proven capable of minimizing the risk of patients with the disease being missed by the diagnosis (false negatives) in order to ensure the effectiveness of medical interventions.

III. RESULTS AND DISCUSSION

Before proceeding to the classification modeling stage, an inferential statistical analysis was conducted on the interrelationships among clinical features using the Pearson Correlation Matrix. This evaluation aimed to map the distribution characteristics of the original data and identify potential multicollinearity issues that could affect the stability of the hybrid model's performance. The detailed results of the correlation coefficient visualization are presented in Figure 2 below:

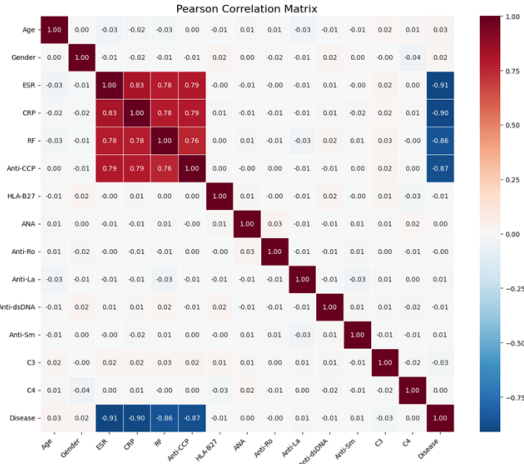


Fig. 2. Feature Correlation Heatmap

Based on the results of the Pearson correlation matrix mapping in Figure 2, a very strong pattern of correlation was found among the clinical variables for rheumatoid arthritis (RA). The main clusters of laboratory tests—namely ESR, CRP, RF, and Anti-CCP—were found to have very high positive linear relationships with one another, as indicated by the dark red areas with correlation coefficients (r) ranging from 0.76 to 0.83. This indicates the presence of multicollinearity, where these independent features contain overlapping (redundant) information because, biologically, they increase simultaneously during the inflammatory phase. On the other hand, these four main features show an extremely strong negative linear relationship with disease status (Disease), indicated by the dark blue areas with strong correlation values ranging from -0.86 to -0.91. Meanwhile, other supporting features such as demographic factors (Age, Gender) and the secondary antibody panel (ANA, HLA-B27, Anti-Ro, Anti-La, Anti-dsDNA, Anti-Sm, C3, C4) are located in the white/pale zone with values close to 0.00, indicating the absence of a direct linear relationship with the patient's diagnostic status in this dataset. The presence of high multicollinearity among the main features (red blocks) is the fundamental reason why the performance of standard classification algorithms may decline if the data is processed directly. These interdependent geometric characteristics of the data provide a strong theoretical foundation and justification for applying the Restricted Boltzmann Machine (RBM) as a latent feature extraction method. RBM works effectively by absorbing information from redundant original features (ESR, CRP, RF, Anti-CCP) and then transforming it into new feature representations in the hidden layer.

A. Data Balancing

In this stage, the SMOTE (Synthetic Minority Over-Sampling Technique) method is used to address data imbalance. SMOTE can also be used to address overfitting. It works by utilizing nearest neighbors (KNN) for each data point in the minority class. The results of the data comparison before and after SMOTE are presented in Figure 3:

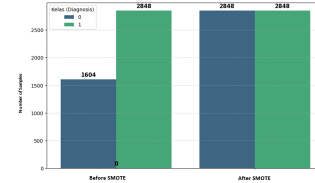


Fig. 3. Target class distribution

Based on the diagram in Figure 3, a comparison of the data distribution before and after applying SMOTE is shown. Before applying SMOTE, the data exhibited class imbalance, where the number of data points in the RA (1) class was significantly higher than in the Normal (0) class. This imbalance can cause the model to be biased toward the RA class (1), thereby reducing its ability to identify the Normal class (0). After applying SMOTE, the number of data points in the Normal class increased significantly through synthetic data generation until it matched the number of RA class data points. This resulted in a more balanced data distribution across classes. Under these conditions, the classification model can learn patterns in both classes more optimally.

B. Hyperparameter Tuning

At this stage, the hyperparameter tuning process uses the GridSearchCV technique to obtain the optimal parameter combination for the RBM model combined with the classification algorithm, using ratios of 70:30, 80:20, and 90:10. The results of the hyperparameter tuning are presented in Table 5:

TABLE V
HYPERPARAMETER TUNING RESULTS

Rasio	70:30	80:20	90:10
RBM-LR			
C	10	10	10
n_components	10	20	20
RBM-RF			
n_components	20	20	20
max_depth	20	20	None
n_estimators	200	100	200
RBM-SVM			
Learning_rate	0.01	0.01	0.01
n_components	10	10	10
C	10	10	10
kernel	rbf	rbf	rbf

Based on the results of hyperparameter tuning in Table V, across the three data splitting scenarios (70:30, 80:20, and 90:10), it is evident that each combination of the hybrid RBM model with a classification algorithm has relatively consistent parameter configurations, with some adjustments to specific parameters to achieve optimal performance. In the RBM-LR model, the constant value of parameter C set to 10 across all ratios indicates that this level of regularization provides a good balance between bias and variance. Meanwhile, the ncomponents parameter increases from 10 at the 70:30 ratio to 20 at the 80:20 and 90:10 ratios, indicating that a greater number of latent features is required to improve data representation as the proportion of training data increases. In the RBM-Random Forest model, the n-components parameter remains

consistent at 20 across all ratios, indicating that the number of features extracted by the RBM is already optimal. The `max_depth` parameter is set to 20 for the 70:30 and 80:20 ratios, but to 'None' for the 90:10 ratio, indicating that an unrestricted tree depth yields better performance when there is more training data. Additionally, the number of `n_estimators` varies between 100 and 200, indicating an adjustment in the number of trees to achieve model stability and accuracy. In the RBM-SVM model, the parameters used tend to remain stable across all ratios, with a learning rate of 0.01, `n_components` of 10, a `C` parameter of 10, and the use of an RBF kernel. This indicates that this configuration is sufficiently optimal for handling non-linear patterns in the data, thus requiring little adjustment even as the data ratio changes. Overall, the tuning results show that an increase in the amount of training data tends to affect the model's complexity requirements, particularly regarding the number of features extracted by the RBM and the parameters in the RF. Meanwhile, the SVM and LR demonstrate parameter stability, indicating that both methods are sufficiently robust to changes in data partitioning.

IV. MODELING AND EVALUATION RESULTS

After performing SMOTE and hyperparameter tuning, the best parameter results will be used to train the models to compare the performance of each classification algorithm based on accuracy, precision, recall, F1-Score, and AUC values in three split ratios, namely 70:30, 80:20, and 90:10. The test results are presented in Table VI.

TABLE VI
MODEL EVALUATION RESULTS

Split Ratio	70:30	80:20	90:10
RBM-LR			
Accuracy	80.75	83.42	79.65
Precision	81.41	83.11	81.47
Recall	81.14	84.96	81.47
F1-Score	81.27	84.02	81.47
AUC	0.8999	0.9193	0.8785
RBM-RF			
Accuracy	93.33	93.95	90.88
Precision	94.31	94.79	93.36
Recall	92.27	93.33	89.78
F1-Score	93.28	94.06	91.53
AUC	0.9822	0.9880	0.9777
RBM-SVM			
Accuracy	99.53	99.56	99.82
Precision	99.21	99.32	99.68
Recall	99.89	99.83	100.00
F1-Score	99.55	99.57	99.84
AUC	0.9999	0.9999	1.0000

Based on the test results for the three data split ratios in Table VI, the RBM-SVM demonstrated the best performance compared to the other models. This is evident from the very high accuracy values: 99.53% for the 70:30 ratio, 99.56% for the 80:20 ratio, and 99.82% for the 90:10 ratio. Additionally, the precision, recall, and F1-Score values are all above 99%, with recall even reaching 100% at the 90:10 ratio. The AUC values, which are close to 1—ranging from 0.9999 to 1.0000—indicate that this model has excellent ability to distinguish between classes, making it the most

optimal model in this study. The RF model ranks second with fairly high and stable performance, with accuracy ranging from 90.88% to 93.95%, peaking at the 80:20 ratio. Precision, recall, and F1-Score values also demonstrate good balance, each exceeding 89%. The AUC values range from 0.9777 to 0.9880, indicating that this model also possesses excellent classification capabilities, although it still falls short of the RBM-SVM's performance. Meanwhile, the RBM-LR model exhibits the lowest performance among the three models, with accuracy ranging from 79.65% to 83.42%. The precision, recall, and F1-Score values range from 81% to 84%, with the best performance at an 80:20 ratio. The AUC values are also relatively low compared to the other models, ranging from 0.8785 to 0.9193, indicating fairly good but not yet optimal classification capabilities. Overall, the results show that the combination of RBM and SVM is capable of producing superior performance in predicting early-stage Rheumatoid Arthritis risk. Additionally, the 80:20 data split ratio tends to yield optimal results for most models, particularly for RBM-LR and RBM-RF. This indicates that the balance between training and test data influences the model's performance in generating accurate and stable predictions.

V. CONCLUSIONS

This study proposes a hybrid framework for the early prediction of rheumatoid arthritis risk by integrating the Restricted Boltzmann Machine (RBM) as a feature extraction method with the Support Vector Machine (SVM), Random Forest (RF), and Logistic Regression (LR) algorithms. The research stages include data preprocessing, Pearson correlation analysis, data balancing using SMOTE, data splitting, hyperparameter tuning using GridSearchCV, latent feature extraction via RBM, and model evaluation using the metrics accuracy, precision, recall, F1-score, and AUC. The results show that RBM is capable of extracting latent patterns and nonlinear relationships in clinical data, thereby improving classification performance. Among the three models tested, the RBM-SVM combination delivered the best performance with an accuracy of 99.82%, a recall of 100%, and an AUC of 1.0000 at a 90:10 data ratio. These findings demonstrate that the integration of RBM and SVM is effective for supporting the early detection of rheumatoid arthritis and minimizing the likelihood of undetected positive cases. This study still has limitations, namely the use of a single public dataset and the comparison of only three classification algorithms without external validation on hospital data. Therefore, future research could develop models using other deep learning architectures, incorporate a wider variety of clinical features, and conduct testing on datasets from various healthcare institutions.

REFERENCES

- [1] M. Ellen M. Gravallese and M. D. Gary S. Firestein, "Rheumatoid Arthritis — Common Origins, Divergent Mechanisms," *N. Engl. J. Med.*, vol. 388, p. 6, 2023, doi: 10.1056/NEJMra2103726.
- [2] S. Alivernini, G. S. Firestein, and I. B. McInnes, "Review: The Pathogenesis of Rheumatoid Arthritis," *Immunity*, vol. 55, no. 12, pp. 2255–2270, 2022, doi: 10.1016/j.immuni.2022.11.009.
- [3] N. I. Butt, B. Waris, and U. Waris, "Delayed Diagnosis and Treatment of Rheumatoid Arthritis: A Preventable Path to Disability," *Am. J. Diagnostic Imaging*, 2025, doi: 10.5455/ajdi.20250531043655.

- [4] S. Amiroch, M. I. Irawan, I. Mukhlash, M. H. Z. Al Faroby, and C. A. Nidom, "Machine Learning for the Prediction of Antiviral Compounds Targeting Avian Influenza A/H9N2 Viral Proteins," 2022, doi: 10.3390/sym14061114.
- [5] A. J. W. Lim et al., "Robust SNP - based prediction of rheumatoid arthritis through machine - learning - optimized polygenic risk score," pp. 1–17, 2023, doi: 10.1186/s12967-023-03939-5.
- [6] J. Jose and D. V. Jose, "Deep learning algorithms for intrusion detection systems in the Internet of Things using the CIC-IDS 2017 dataset," *Int. J. Electr. Comput. Eng.*, vol. 13, no. 1, pp. 1134–1141, 2023, doi: 10.11591/ijece.v13i1.pp1134-1141.
- [7] J. Jamil and S. Pulukadang, "Application of Deep Learning Methods in Education," vol. 4, no. 6, pp. 1019–1038, 2025.
- [8] M. Javeed, M. Gochoo, A. Jalal, and K. Kim, "HF-SPHR: Hybrid Features for Sustainable Physical Healthcare Pattern Recognition Using Deep Belief Networks," 2021, doi: 10.3390/su13041699.
- [9] Yudicy Amelia, Puspa Eosina, Foni Agus Setiawan, "Comparison of Deep Learning and Machine Learning Methods for Classification (Trial on Breast Cancer Data)," pp. 109–114, 2018, DOI: 10.32832/inovatif.v1i2.544
- [10] R. R. Salam, M. F. Jamil, and Y. Ibrahim, "Sentiment Analysis of Cash Direct Assistance Distribution for Fuel Oil Using Support Vector Machines," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 3, no. April, pp. 27–35, 2023.
- [11] G. Li, A. Zhang, Q. Zhang, D. Wu, and C. Zhan, "Pearson Correlation Coefficient-Based Performance Enhancement of a Broad Learning System for Stock Price Prediction," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 69, no. 5, pp. 2413–2417, May 2022, doi: 10.1109/TCSII.2022.3160266," *Inst. of Electr. and Electron. Eng.*, vol. 69, no. 5, pp. 2413–2417, 2022, doi: 10.1109/TCSII.2022.3160266.
- [12] Y. Zhang, L. Deng, and B. Wei, "Imbalanced Data Classification Based on Improved Random-SMOTE and Feature Standard Deviation," *MDPI*, vol. 12, no. 11, pp. 1–17, 2024, doi: <https://doi.org/10.3390/math12111709>.
- [13] V. R. Joseph, V. R. Joseph, and H. M. Stewart, "Optimal ratio for data splitting," *Stat. Anal. Data Min. ASA Data Sci. J.*, vol. 15, no. 4, pp. 531–538, 2022, doi: 10.1002/sam.11583.
- [14] M. Sivakumar and S. Parthasarathy, "Trade-off between training and testing ratio in machine learning for medical image processing," pp. 1–17, 2024, doi: 10.7717/peerj-cs.2245.
- [15] A. Setiawan, A. A. Siregar, N. Setiawan, J. Nasution, and N. D. Putra, "Optimization of SVM and Random Forest Model Performance for Breast Cancer Classification Using Hyperparameter Tuning," vol. 4, no. 3, pp. 2141–2149, 2026, doi: 10.62712/juktisi.v4i3.789
- [16] B. Santoso, "Unsupervised Feature Extraction Using Restricted Boltzmann Machines: A Probabilistic Approach in Deep Learning Applications," *Journal of Informatics*, vol. 18, pp. 45–58, 2023.
- [17] J. Schlüter, "Restricted Boltzmann Machine Derivations," no. 6, pp. 1–18, 2014.
- [18] S. N. Khan, S. U. Khan, H. Aznaoui, and C. B. Şahin, "Generalization of linear and non-linear support vector machines in multiple fields: a review," vol. 4, no. 3, pp. 226–239, 2023, doi: 10.11591/csit.v4i3.pp226-239.
- [19] E. Najjar, N. A. Alkhaykanee, and A. Majeed, "Classification of Arabic Documents with Five Classifier Models Using Machine Learning," vol. 9, no. January, pp. 365–370, 2025.
- [20] N. Alexander, R. Bria, and A. Witanti, "Analysis of Indonesian Public Sentiment Using the Support Vector Machine Algorithm Regarding the 2024 Presidential Election," vol. 7, no. 6, 2024.