

Soft-Voting Ensemble Learning for Multimodal Rheumatoid Arthritis Classification

Sharik Ahmad* and Rafat Parveen

Department of Computer Science, Jamia Millia Islamia, New Delhi, India – 110025

Received: 3 Jun. 2026, Revised: 19 Aug. 2026, Accepted: 23 Aug. 2026

Published online: 1 Sep. 2026

Abstract: This article presents a multimodal machine learning (M3L) framework for the early prediction and classification of rheumatoid arthritis (RA) severity using a combination of clinical and biomedical data. Current approaches frequently rely on single models or unimodal datasets, which may reduce both accuracy and robustness. To overcome this challenge, the study introduces an ensemble learning method using a soft-voting classifier that integrates Gradient Boosting, Random Forest, and XGBoost. The proposed model determines RA severity stages based on informative features such as ESR, CRP, anti-CCP antibody levels, MRI joint inflammation indicators, gene expression scores, and additional relevant variables. By leveraging multiple algorithms, the framework mitigates overfitting and improves generalization capability. Performance is evaluated using accuracy, precision, recall, and F1-score, supported by comparative visual analyses to examine each model's strengths and weaknesses. Results demonstrate that the ensemble achieves an accuracy of 0.88, outperforming XGBoost (0.86), Random Forest (0.86), and Gradient Boosting (0.87).

Keywords: Rheumatoid Arthritis, Machine Learning, Multimodal biomedical data, Voting Classifier, XGBoost, Random Forest, Gradient Boosting

1 Introduction

Rheumatoid arthritis (RA) is a chronic autoimmune disease that causes inflammation of the joints, causing pain, swelling and eventual joint deformity. Early detection and accurate classification of the severity of RA are important for appropriate treatment and management of the disease. Precision and predictive accuracy in the diagnostic methods is often limited to clinical assessments, biomarker analysis and imaging [1]. Therefore, data-driven machine learning (ML), and more specifically, ensemble learning methods, can be used for improving RA diagnosis. These are some of the methods, among which the Voting Classifier is an effective ensemble learning algorithm that increases prediction accuracy by combining other models.

The Voting Classifier is a very powerful Machine learning technique; it takes multiple classifiers' predictions and consolidates the entire prediction to make a final decision. Unlike the individual models, these bias or variance issues affecting individual models, the ensemble method becomes harder to be biased and improves generalization and robustness [2]. The Voting Classifier is integrated with three ML models, namely XGBoost, Random Forest, and Gradient Boost, to predict the RA severity in this research. Ran-

dom Forest is efficient in learning from structured medical data and preventing overfitting, XGBoost is good at learning from high-dimensional data and finding out complex patterns, and Gradient Boosting is capable of improving accuracy through iterative learning.

The Voting Classifier exploits the joint power of these classifiers to improve prediction reliability and is well-suited for medical applications. Across all cases, moreover, no single model is optimal in medical diagnostics. With the Voting Classifier, in order to mitigate this problem, they either hard vote, giving the final outcome according to the majority, or they soft vote, where they average the class probabilities from all models and make the final decision. Soft Voting is used in this research as it gives greater variation in the classification than the use of one model [3]. As RA progression is a complex process for a number of biological and clinical factors, a soft voting ensemble balances the decision-making impact on misclassified cases from any one model.

The key advantage of this approach lies in the capacity to incorporate multi-modal medical data comprising clinical biomarkers of Erythrocyte Sedimentation Rate (ESR), C-reactive Protein (CRP), anti-CCP antibodies, imaging

* Corresponding author e-mail: shariqgee@gmail.com

features like MRI joint inflammation and Multimodal biomedical data. Such complex and heterogeneous datasets are problematic for traditional statistical tools but are an easier problem for machine learning, especially when the ensemble model is involved in capturing the intricate relationship between various variables. This study approaches conventional RA diagnostic methods using advanced data-driven techniques that enhance the accuracy of classification and incorporate Multimodal biomedical data.

The aim of this research is to design a Multi-modal voting classifier-based framework that cannot just predict RA, but also predict its severity levels, thus aiding in designing the personalized treatment strategy. The proposed framework provides a scalable, interpretable and clinically useful model, which can facilitate medical practitioners to detect early RA and make decisions using a model. This study shows that ensemble learning has advantages in healthcare by comparing the Voting Classifier's performance with that of its constituent models. Multiple machine learning techniques together minimise the risk of overfitting and generalise the model, thus improving patient outcomes.

2 Literature Review

2.1 Machine Learning in Rheumatoid Arthritis Diagnosis

Machine learning (ML) has led the way to a transformative approach to medical diagnostics, especially for the complexity and heterogeneity of Rheumatoid Arthritis (RA) disease. RA is an autoimmune disorder in which pain, swelling, and disability result from the breakdown of the joint. Early detection and sensitive discrimination of RA severity are important for effective disease management and the design of effective treatment plans [4]. However, traditional methods of diagnosis based on physical markers (ESR, CRP, anti-CCP antibodies), radiographic imaging and physician metrics have traditionally been subjective and subject to delay. The challenges here have been overcome through the use of machine learning techniques that increased both RA diagnosis and classification automation and enhancement.

Throughout the years, many types of ML models have been used to analyse structured and unstructured clinical data used to including Support Vector Machine (SVM), Decision Trees, and Artificial Neural Network (ANN). In particular, these models will be able to identify the hidden patterns in large datasets that one may not notice with traditional statistical methods [4]. For example, a number of studies show that SVM is able to distinguish RA patients and healthy individuals on the basis of biomarkers and imaging data. Similarly, Decision Trees are an interpretable classification framework that is useful for medical decision making. Deep learning architectures have had remarkable success at the medical imaging tasks, particularly in tasks associated with the identification of joint inflammation in MRI scans [5]. However, as with many modern application

areas of ML, a major challenge in applying ML to RA diagnosis is the high dimensionality of biomedical data.

Often, given the complex relationships between clinical, molecular and imaging variables, a single ML model has limited ability to achieve consistent classification performance. In order to overcome these problems, ensemble learning techniques have been used to augment diagnoses of prostate cancer with multiple models, enhancing diagnostic accuracy and robustness. The heterogeneity of RA data lends itself well to the reduction of variance and the mitigation of overfitting by ensemble methods, which improves predictive performance. These techniques lead to significant improvements in RA classification among these techniques, Voting Classifiers, Bagging and Boosting methods (such as Random Forest, XGBoost and Gradient Boosting). In these models, final classifications are made by aggregating predictions from a collection of base learners, in order to obtain generalization.

Voting Classifier is one of the best ensemble solutions for a RA diagnosis by integrating multiple algorithms such as Random Forest, XGBoost and Gradient Boosting to finally get more reliable results. The soft voting strategy is applied in this classifier through averaging of class probabilities of different classifiers, which creates a more balanced classification with respect to the RA severity levels. It has been observed that Voting Classifiers outperform individual ML models in the classification of RA patients using clinical and molecular features [6]. Tree-based models such as Random Forest and Gradient Boosting increase feature selection, and XGBoost's optimal learning mechanism overcomes this problem permanently [7].

Apart from classification, ML models are also being used to prognosticate RA based on historical patient data. Using multi-modal data sets such as clinical, imaging and biomedical data can integrate ML to gain a more comprehensive view of RA mechanisms. Now, as more research is conducted, ML is applied to the RA diagnosis application, and it establishes a data-driven, precision medicine approach to improve patient outcomes using early detection and personalized diagnosis and treatment strategies.

2.2 Ensemble Learning for Medical Classification Problems

In medical classification problems where disease classification has to be done early and accurately, especially that of Rheumatoid Arthritis (RA), ensemble learning has become and still is an important technique. Ensemble learning offers a significant advantage of taking advantage of combining multiple models to improve the prediction performance as well as to reduce bias, thereby also reducing overfitting. However, among all ensemble techniques, Voting Classifiers, Bagging and Boosting have proven to be very successful in dealing with complex medical datasets. These methods combine predictions from different base learners to get greater classification accuracy than that of a single model [8].

Multiple ensemble approaches have been tried in the context of RA diagnosis and RA severity prediction. As these are the most widely used models, they use Random Forest, XGBoost, and Gradient Boosting to respond to the noisy, high-dimensional medical data. As a bagging-based algorithm, Random Forest builds multiple decision trees on different subsamples of the data and averages their predictions are robust to the noise and overfitting. In contrast, XGBoost and Gradient Boosting make use of an iterative boosting process where weak learners are sequentially improved in order to achieve non-linear relationships as well as intricate disease patterns. An application of these models has been applied to predict RA severity, treatment response, and patient stratification using clinical and molecular biomarkers.

Although each of Random Forest, XGBoost and Gradient Boosting is effective, there are limitations with each of them. Although robust to noise, Random Forest does not have the ability to learn deep feature interactions found in medical data. While powerful, both XGBoost and Gradient Boosting are very sensitive to hyperparameter tuning and are highly susceptible to overfitting if not well regularized [9]. To address each of these handicaps, researchers have more and more turned to the Voting Classifier, an ensemble technique that enables multiple algorithms to construct a better balanced and expansive classification model.

The Voting Classifier is a method of averaging predictions from a number of base learners, e.g., Random Forest, XGBoost, Gradient Boosting, and making an inference by voting majority (hard voting) or averaging probabilities (soft voting). In medical classification, the soft voting strategy is very helpful because it considers the confidence level of each model and makes a more stable and accurate prediction. Integration of these models within the Voting Classifier framework has been demonstrated to have a dramatic effect on the accuracy of classification and patient stratification, and severity of disease predictions as compared to individual models.

A major advantage of the Voting Classifier is that it takes the strengths of different ML algorithms. Random Forest provides robustness and feature selection, XGBoost improves predictive performance because of better boosted decision trees, and Gradient Boosting supports fine-tuned learning of complex data relationships. Due to these strengths, the Voting classifier helps mitigate overfitting, balances the bias-variance trade-off and generalizes better in unseen medical datasets.

The Voting Classifier has also been used successfully in other domains in medical areas, such as the prediction of neurological disorders, cancer detection, and cardiovascular disease. The potential of ensemble learning techniques in healthcare analytics for improving early disease detection, enhancing treatment planning and advocating precision medicine initiatives is being realized with the growing adoption of ensemble learning techniques [10]. However, with research advancing, medical classification accuracy and decision-making are predicted to be further refined

Table 1: Multi-Modal Data Sources for RA Prediction

Data Type	Key Features	Diagnostic Contribution
Clinical Biomarkers	ESR, CRP, Anti-CCP, RF Factor	Inflammation & immune response
Imaging Data	MRI joint scores, Ultrasound	Joint erosion & synovitis detection
Molecular Data	Cytokines, Proteomic profiles	Immune dysregulation tracking
Genetic Data	HLA-DR4, SNP Variants	Genetic predisposition to RA

by hybrid ensemble approaches using deep learning and traditional ML models.

2.3 Multi-Modal Data Integration in RA Prediction

Rheumatoid arthritis (RA) is a complicated, heterogeneous, variable disease with regard to its progression and biomarker profiles. It is a difficult problem for its very nature, and the conventional diagnostic approaches based only on clinical observations and the patient's reported symptoms cannot always reach the high precision of the classification. To address this, various researchers have focused on multi-modal data integration, i.e., clinical, imaging, molecular, and genetic biomarkers integration to assist RA diagnosis and severity prediction.

Machine learning (ML) models can benefit from integrating multimodal data, which enables ML models can utilize more disease indicators, reduce diagnostic uncertainty, and thus improve classification performance [11]. In RA, Clinical biomarkers that are Erythrocyte Sedimentation Rate (ESR), C Reactive Protein (CRP) and Anti Cyclic Citrullinated Peptide (Anti CCP) Antibodies are important to diagnose but may not be specific enough for Early RA detection. MRI and ultrasound are used to study the joint inflammation and structural damage and monitor disease progression. Furthermore, genetic and molecular markers give information (Table 1) about the underlying pathophysiological mechanisms of RA and help in personalized treatment strategies.

RA prediction is improved by integrating these data types into ML-driven ensemble models, for example, Voting Classifiers, doing so combines different levels of information about the disease. For example, deep learning based on imaging can extract features from MRI scans, and basic ML models can analyze clinical or molecular markers. The combination of these approaches results in a more comprehensive RA classification system that should improve early disease detection, severity grading, and prognosis of treatment response.

In addition, recently, there have been studies of the use of biomedical and molecular biology-driven datasets to get deep into the biology of RA. For instance, single-cell RNA sequencing has its breakthrough in identifying specialized immune cell subtypes that advance the progression of RA. Both metabolomics and lipidomics data can be used to

distinguish RA patients from non-RA individuals by revealing specific metabolic signatures [12]. By integrating these high-dimensional datasets into the ensemble learning, models become robust and have higher predictive accuracy.

One of the unique challenges in the task of multi-modal integration is heterogeneity and dimensionalization of the medical data. It also means ensuring that missing data does not affect the quality of the model, and making sure the data types are normalised if they are in a variety of formats, as well as finding the meaningful patterns across multiple modalities. For example, Feature selection techniques, RFE and PCA are used frequently for optimizing the ML models' prediction of multi-modal RA.

Deep learning-based fusion techniques for further improving multimodal data fusion are being studied as the research progresses. The next generation of precision medicine approaches for RA diagnosis and management should involve hybrid models that combine Convolutional Neural Networks (CNNs) for imaging data with ensemble classifiers for clinical/molecular data.

2.4 Recent Advances in Pattern Recognition for Multimodal Medical Data Analysis

With the recent advances in pattern recognition and machine learning, the detection of diseases using a combination of diverse biomedical data has become more precise. Deep learning and ensemble methods are used to identify the complex patterns hidden in biomedical data. For example, Roy et al. proposed a novel ensemble learning method that increases the accuracy of medical image classification using a combination of deep learning methods and diversity control during the training period [13]. Additionally, Karthik et al. proposed a multimodal approach that combines medical image data with deep learning and ensemble methods to classify medical images and demonstrated the effectiveness of this integration in improving disease classification precision. Recently, Kumar et al. proposed a novel method using ensemble learning and machine learning to accurately predict medical diseases. Furthermore, Kumar and Dahiya introduced a segmentation-driven ensemble learning framework that integrates feature extraction and ensemble classifiers to improve classification accuracy in medical imaging tasks. Inspired by these developments, this study proposes a multimodal ensemble framework that integrates tree-based classifiers to improve the prediction of rheumatoid arthritis severity.

2.5 Performance Evaluation Metrics in Medical ML Models

To determine the efficacy of ML models for Rheumatoid Arthritis (RA) classification, it is vital to employ robust evaluation metrics to achieve high accuracy and clinical relevance. Different metrics are used in healthcare AI research

to characterize the predictive capacity of models for the task of distinguishing RA from non-RA cases, for example, or to distinguish RA severity levels. The most common metrics used to evaluate are Accuracy, Precision, Recall, F1 score and Area under the curve - Receiver Operating Characteristic (AUC-ROC). These metrics help to understand and evaluate how well the model performs and whether or not it can make clinically meaningful predictions.

Accuracy: The accuracy metric is the simplest, which is the proportion of correctly classified cases. While it is reasonable when we have balanced datasets, it isn't precise for the imbalanced datasets, such as the presence of RA cases could be a lot fewer than the non-RA cases [14].

Precision (Positive Predictive Value - PPV): Precision represents the ratio of correctly predicted True RA cases to all predicted RA cases. Minimizing false positives is of utmost importance because classifying a non-RA patient as having the disease would be critical.

Recall (Sensitivity or True Positive Rate - TPR): The aim of recall is to find out what proportion of actual RA cases the model identified. As a high recall score helps in preventing false negatives, RA patients are to be diagnosed and treated at the earliest.

F1-Score: The harmonic mean of Precision and Recall, known as F1-score, is the balanced metric between Precision and Recall. When dealing with an imbalanced dataset, it is especially helpful in ensuring low false positives and false negatives.

AUC-ROC: The AUC-ROC metric measures the model's capability of distinguishing between cases with RA and without RA at different classification thresholds. The higher the AUC value, the better the model performs.

Ensemble vs. Individual Models: Previous studies have revealed that ensemble learning models are far more powerful in RA diagnosis compared to individual linear classifiers. Ensemble techniques, by combining multiple model predictions, increase the AUC-ROC score and balanced F1-score over individual models while improving reliability for clinical applications [15].

3 Methodology

3.1 Research Approach

The research uses quantitative data methods with supervised machine learning algorithms to develop a predic-

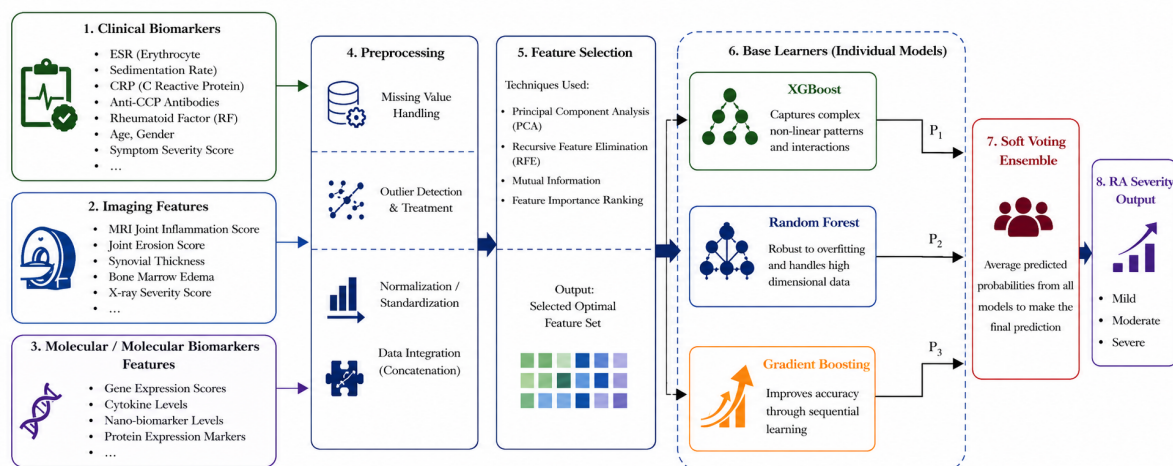


Fig. 1: Overview of the Multi-Modal Research Approach Flowchart

tive model for Rheumatoid Arthritis (RA) [16]. The selected method enables efficient operation on biomedical data, which includes clinical biomarkers together with nano-informatics indicators alongside genetic expressions. Standard statistical methods cannot detect difficult data relationships in the research because these contain both structured and unstructured biomedical signals. The research aims to deliver two essential outcomes by using clinical indicator examination for early detection combined with a technology that combines nano-informatics information with molecular and cellular evaluation systems for precise patient status measurement. The research depends on Voting Classifier ensemble learning to integrate Random Forest, XGBoost, and Gradient Boosting predictive models for its goal achievement. Supervised learning represents a suitable approach because the dataset contains specific outcome labels that define RA severity levels [17]. Multiple markers of inflammation (ESR, CRP), joint degradation values, genetic background information, and nanoparticle response characteristics feed into model training because these represent an extensive awareness of RA biological processes:

$$f(x) \rightarrow y \quad (1)$$

$$f = \arg \min_{f'} \sum_{i=1}^n L(f'(x^{(i)}), y^{(i)}) \quad (2)$$

Python implementation snippet for data ingestion:

```
import pandas as pd
df = pd.read_csv('RA_Dataset.csv')
print(df.head())
print(df.info())
```

3.2 Research Design

The research utilizes an experimental design structure because different machine learning models receive training

and validation before teams perform model benchmarking through their performance metrics. Model development starts after data acquisition, preprocessing, feature selection, and evaluation stages.

Data Preprocessing: The cleaned dataset includes duplicate removal coupled with value handling methods such as imputation or exclusion, along with field normalization for numeric values. Categorical data is converted using one-hot encoding or label encoding [18]. To handle high-dimensional biomedical data, PCA and t-SNE are used for dimensionality reduction without losing critical information variance.

Feature Selection and Engineering: The most predictive variables are identified through feature importance metrics using tree-based models in collaboration with rheumatology and molecular biology experts. Feature engineering includes polynomial expansions, interaction terms, and signal smoothing.

Model Development: Three main classifiers—Random Forest, XGBoost, and Gradient Boosting—are trained independently. A Voting Classifier combines these models using a soft voting strategy averaging class probabilities for optimal accuracy [19]. Hyperparameters are tuned via grid search and cross-validation.

Model Evaluation: Key performance indicators consist of precision, accuracy, recall, F1-score, and AUC-ROC, alongside confusion matrices for in-depth error analysis.

```
# Data Preparation & Splitting
df['Gender'] = df['Gender'].map(
```

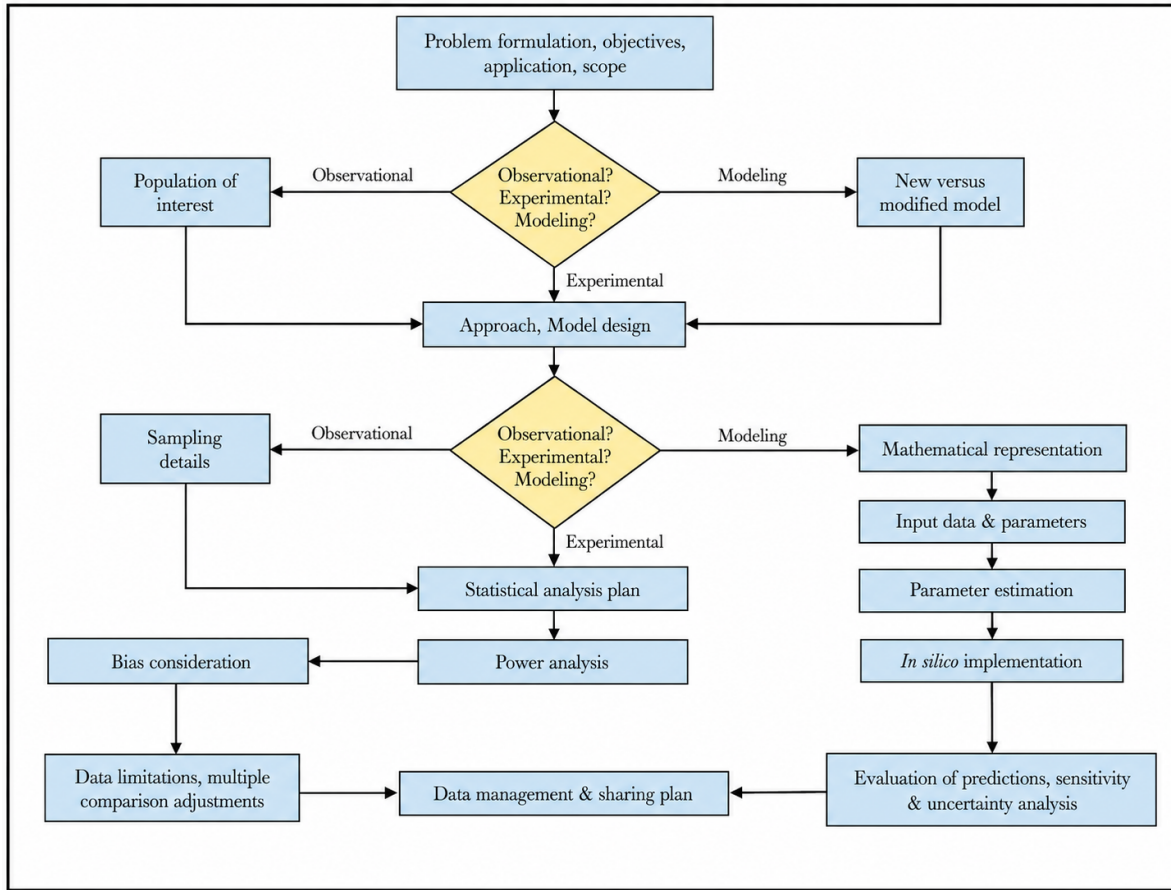


Fig. 2: Experimental Design and Model Development Framework Flowchart

```

{'Male': 0, 'Female': 1})
df['Anti_CCP_Antibodies'] = \
    df['Anti_CCP_Antibodies'].map(
        {'Positive': 1, 'Negative': 0})
X = df.drop(columns=['Patient_ID',
                    'RA_Severity'])
y = df['RA_Severity']
from sklearn.model_selection \
    import train_test_split
X_tr, X_te, y_tr, y_te = \
    train_test_split(
        X, y, test_size=0.3,
        random_state=42)

```

$$r_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} \quad (3)$$

Algorithm 1: Proposed Ensemble Voting Algorithm for RA Severity Prediction

–Input: Dataset D , Feature matrix $X = \{x_1, \dots, x_n\}$, Target vector y , Fold count $k = 5$, Base models $M = \{M_1 : \text{XGBoost}, M_2 : \text{Random Forest}, M_3 : \text{Gradient Boosting}\}$, Voting mechanism V (Soft Voting).

–Output: Trained Voting Classifier V_{trained} , Predictions $\hat{y}_{\text{pred}}$, Evaluation metrics.

Procedure:

- 1.Preprocessing:** Clean missing values, scale continuous features, encode categorical, split into D_{train} (80%) and D_{test} (20%).
- 2.Model Initialization:** Set hyperparameter search grids for M_1, M_2, M_3 .
- 3.Tuning:** Execute 5-fold cross-validation grid search on D_{train} and refit optimal estimators.
- 4.Ensemble Construction:** Instantiate V as a soft-voting classifier combining M_1, M_2, M_3 ; train V on D_{train} .
- 5.Evaluation:** Predict $\hat{y}_{\text{pred}} = V.\text{predict}(X_{\text{test}})$; compute Accuracy, Precision, Recall, F1, and AUC-ROC.
- 6.Clinical Insights:** Extract and validate top feature importances against clinical benchmarks.

Complexity: Training: $O(T_{\text{XGB}} + T_{\text{RF}} + T_{\text{GB}})$; Prediction: $O(n \cdot (M_1 + M_2 + M_3))$; Space: $O(n \cdot d + M_{\text{params}})$.

Table 2: Sample Patient Data Records

Feature	P0055	P0115	P0112	P0079	P0147
Age	20	38	74	78	50
Gender	Female	Male	Female	Male	Male
Symptom Severity	10	7	4	10	5
ESR (mm/hr)	64.4	32.0	75.8	22.5	78.1
CRP (mg/L)	15.0	26.2	45.8	43.5	12.7
Anti-CCP Antibodies	Negative	Positive	Positive	Negative	Positive
RF Level (IU/mL)	65.8	65.0	89.6	15.4	48.5
X-Ray Severity Score	0	0	1	4	0
MRI Joint Inflammation	24.7	51.3	65.1	26.0	53.3
Nano Biomarker Level	5.76	4.28	5.39	9.56	1.51

3.3 Data Collection and Analysis

This study uses a dataset obtained from Kaggle containing clinical, imaging, and molecular biomarker data for Rheumatoid Arthritis diagnosis. Patient records include ESR, CRP, anti-CCP antibodies, MRI joint inflammation scores, and genetic markers [20]. The dataset was preprocessed through mean/mode imputation, normalization, and one-hot encoding [21].

All data preprocessing, model training, and evaluation took place in Python using Pandas, NumPy, Scikit-learn, and Matplotlib.

4 Model Development in Multimodal Machine Learning

Table 2 displays representative patient records from the dataset, illustrating features including Patient.ID, Age, Gender, Symptom Severity, ESR, CRP, Anti-CCP Antibodies, RF Level, X-Ray Severity Score, MRI Joint Inflammation, and Nano Biomarker Level [1].

4.1 Exploratory Data Analysis (EDA)

Table 3 summarizes the statistical distribution of all features in the dataset ($N = 363$), including mean, standard deviation, minimum, quartiles, and maximum values.

Table 4 presents the age-wise distribution of RA severity levels across patient age groups. Patients in the 70–80 age group exhibit the highest proportion of high-severity RA cases (48.78%), demonstrating that disease severity tends to increase with age. In contrast, the 50–60 and 60–70 groups have higher proportions of low-severity cases.

The boxplot in Fig. 3 illustrates the distribution of ESR and CRP inflammation markers. The values of ESR demonstrate higher variability across arthritic patients compared to CRP. The age distribution reveals a higher concentration of patients in the 70+ demographic, affirming age as a primary risk factor.

Fig. 4 compares Symptom Severity across RA Severity levels, showing that higher RA severity corresponds to higher symptom scores.

Fig. 5 displays the pairwise feature correlation heatmap, highlighting a notable correlation between Gene Expression Score and Symptom Severity [22].

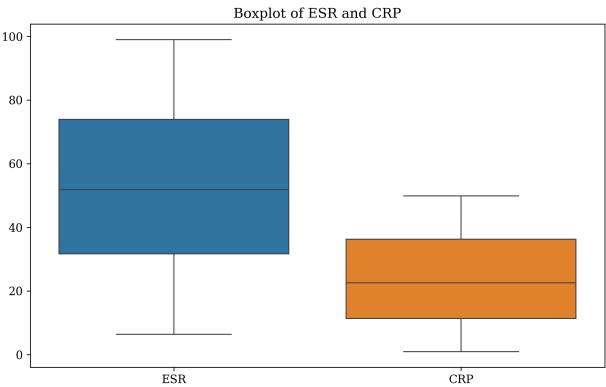


Fig. 3: Boxplot of ESR and CRP Distribution for Arthritic Patients

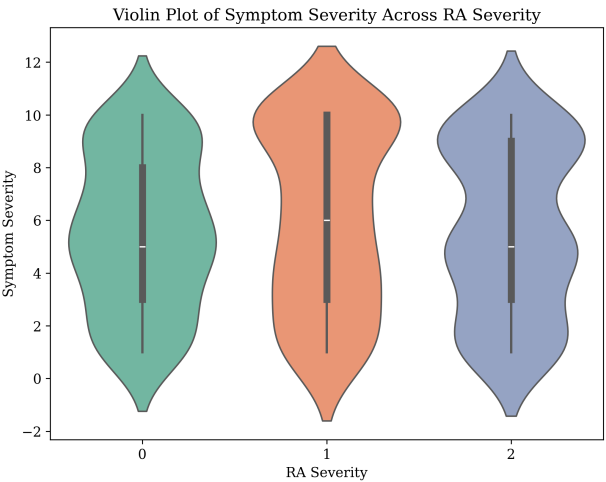


Fig. 4: Violin Plot of Symptom Severity Across RA Severity Levels

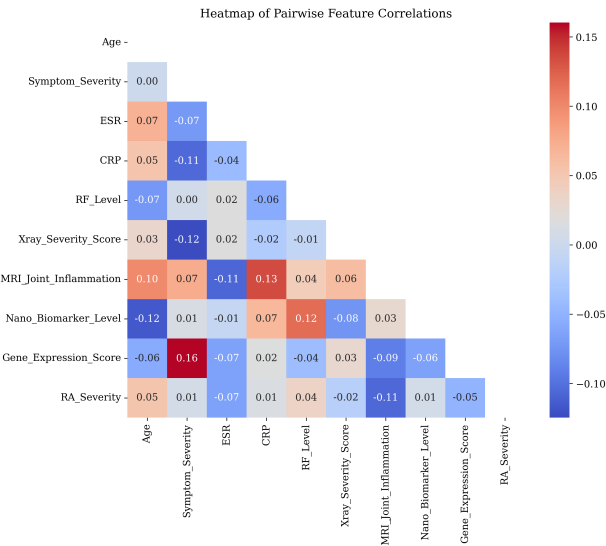


Fig. 5: Feature Correlation Heatmap

Table 3: Statistical Summary of Dataset Features ($N = 363$)

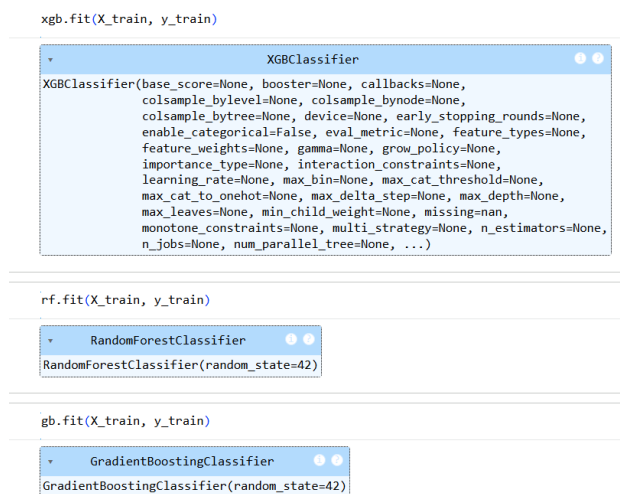
Metric	Age	Symptom Severity	ESR	CRP	Anti-CCP Antibodies	RF Level	Joint Swelling	MRI Results	Gene Expr. Score	RA Severity
Count	363	363	363	363	363	363	363	363	363	363
Mean	50.77	52.28	23.83	53.53	2.47	48.42	5.018	49.33	1.000	1.000
Std Dev	18.56	25.24	14.50	27.03	1.654	29.98	2.880	29.58	0.817	0.817
Min	18.00	6.40	1.00	10.50	0	1.20	0.120	0.40	0	0
25%	33.50	31.65	11.35	29.10	1	21.30	2.960	24.20	0	0
50%	52.00	51.80	22.60	53.50	3	46.40	5.160	49.90	1	1
75%	68.00	73.90	36.30	78.60	4	76.60	7.590	76.45	2	2
Max	79.00	99.00	49.90	99.70	5	99.80	9.980	99.50	2	2

Table 4: Age-wise RA Severity Distribution

Age Range	Low Severity (%)	Mild Severity (%)	High Severity (%)
50–60	45.71	35.71	18.57
60–70	42.50	37.50	20.00
70–80	24.39	26.83	48.78

Table 6: Cross-Validation Performance Comparison

Model	Test Accuracy	Mean CV Accuracy	Std Dev
XGBoost	0.86	0.85 ± 0.03	0.03
Random Forest	0.86	0.84 ± 0.04	0.04
Gradient Boosting	0.87	0.86 ± 0.03	0.03
Voting Classifier	0.88	0.87 ± 0.02	0.02

**Fig. 6:** Sub-Model Fitting Workflow**Table 5:** Classification Performance Report for Evaluated Models

Metric	XGBoost	Random Forest	Voting Classifier	Gradient Boosting
Accuracy	0.86	0.86	0.88	0.87
Macro Precision	0.87	0.87	0.88	0.88
Macro Recall	0.86	0.86	0.89	0.88
Macro F1-Score	0.86	0.87	0.88	0.87
Weighted Precision	0.87	0.87	0.89	0.88
Weighted Recall	0.86	0.86	0.88	0.87
Weighted F1-Score	0.86	0.86	0.88	0.87

4.2 Model Development and Comparative Performance

Fig. 6 illustrates the independent training process of XGBoost, Random Forest, and Gradient Boosting models on the training dataset [23].

Table 5 presents the full classification report comparing XGBoost, Random Forest, Gradient Boosting, and the Voting Classifier across precision, recall, and F1-score metrics.

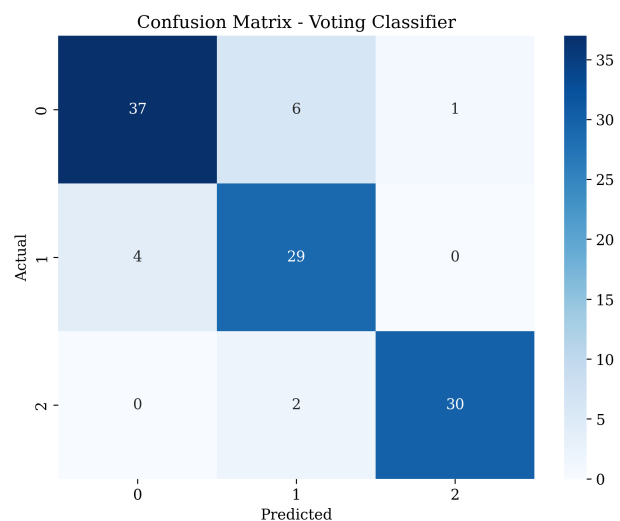
**Fig. 7:** Confusion Matrix for the Proposed Soft-Voting Classifier

Table 6 compares individual models with the proposed soft-voting ensemble across 5-fold cross-validation. The Voting Classifier achieved the highest accuracy (0.88) and mean cross-validation accuracy (0.87 ± 0.02), with the lowest standard deviation indicating superior stability.

Fig. 7 shows the confusion matrix for the Voting Classifier on the test dataset across RA severity classes (0: Low, 1: Mild, 2: High).

Fig. 8 provides a comparative bar chart of accuracy, precision, recall, and F1-score across all models [24], demonstrating the clear advantage of the soft-voting ensemble.

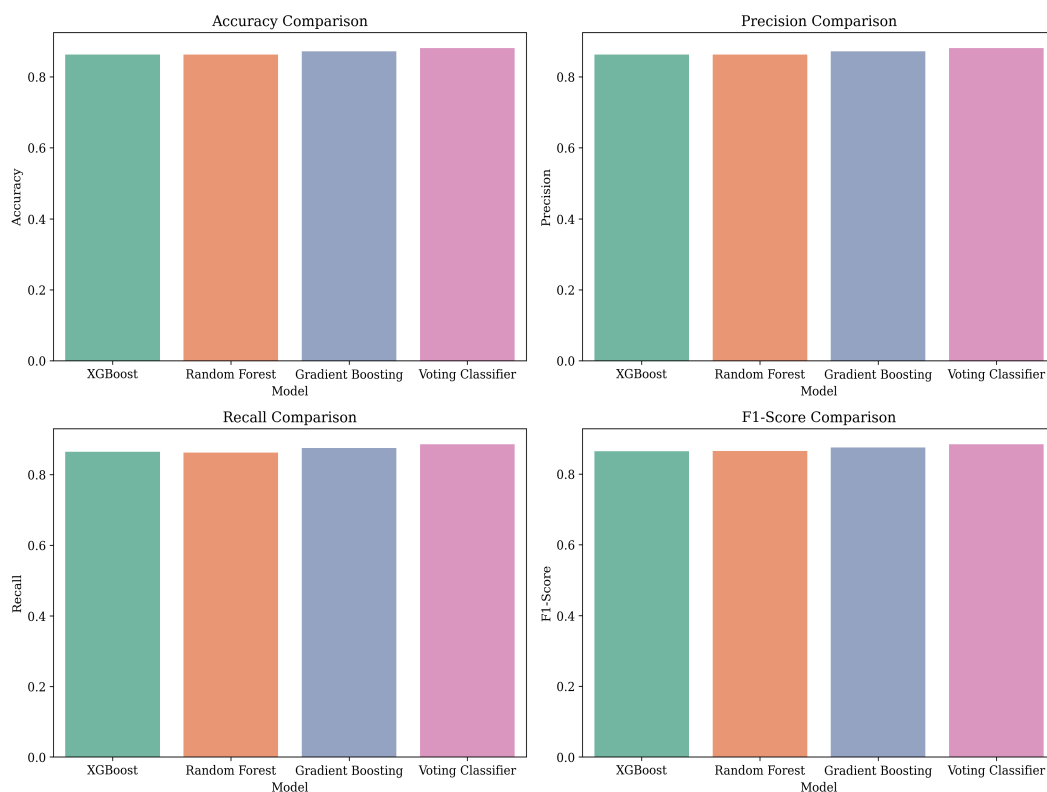


Fig. 8: Evaluation Metric Comparison Across All Models

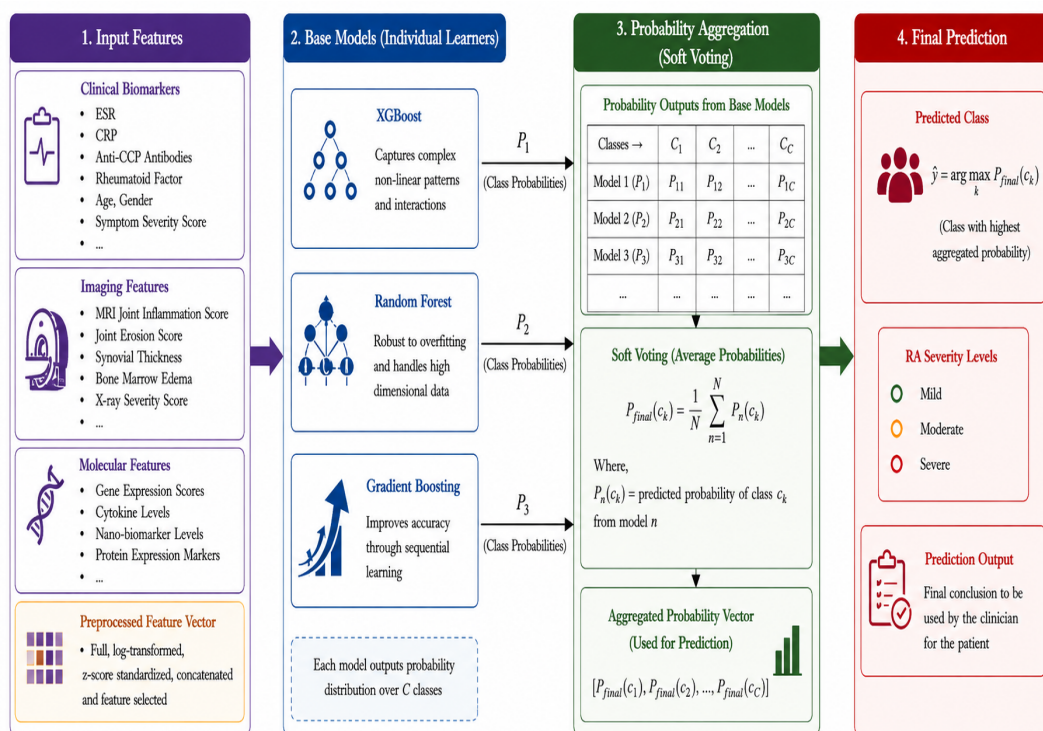


Fig. 9: Complete Soft-Voting Ensemble Workflow Architecture

4.3 Novelty and Workflow Architecture

The novelty of this study lies in integrating complementary tree-based learners (XGBoost, Random Forest, Gradient Boosting) within a soft-voting ensemble framework tailored for multimodal RA severity classification (Fig. 9) [25].

5 Discussion

A multimodal machine learning framework was presented for RA severity prediction from structured clinical, imaging, and biomarker data [26]. The results underscore the critical importance of multimodal integration in enhancing diagnostic accuracy and clinical decision support.

Data Analysis Insights

The dataset encompassed Patient ID, Age, Gender, Symptom Severity, ESR, CRP, Anti-CCP Antibodies, RF Level, X-Ray Score, MRI Joint Inflammation, and Nano-Biomarker levels. ESR values showed wider dispersion than CRP, suggesting ESR is a highly sensitive indicator of patient-specific disease variation [27]. Furthermore, age distribution analysis confirmed elevated RA prevalence among patients aged 70 and above, consistent with medical literature. Symptom severity strongly correlated with higher RA severity classes, and moderate correlation with Gene Expression Scores highlights the indispensable role of genetic markers in disease tracking.

Machine Learning Performance

The soft-voting ensemble achieved superior performance across all evaluation metrics compared to individual base learners [28]. The ensemble's probability averaging effectively mitigated the individual biases and variances of single models, resulting in reduced misclassification and enhanced generalization stability.

Clinical Implications and Limitations

From a clinical standpoint, this framework serves as a proof of concept for AI-driven multi-modal diagnostics in rheumatology, aiding clinicians in early detection and personalized intervention.

Limitations: Key limitations include reliance on a secondary dataset of moderate size, absence of external prospective clinical validation, lack of longitudinal follow-up records, and potential sensitivity to imputation assumptions. Future research will explore deep multimodal fusion architectures, wearable IoT continuous monitoring, and Explainable AI (XAI) tools such as SHAP and LIME to enhance clinical trust [29].

6 Conclusion

This study demonstrated that a Soft-Voting Ensemble Classifier integrating XGBoost, Random Forest, and Gradient Boosting significantly enhances Rheumatoid Arthritis severity prediction using multimodal clinical, imaging, and molecular biomarker data. The ensemble achieved 0.88 accuracy and 0.87 ± 0.02 cross-validation accuracy, outperforming all individual classifiers. Future efforts will focus on expanding dataset diversity, incorporating real-time wearable monitoring, and embedding Explainable AI frameworks for clinical deployment.

Declarations

Competing Interests: The authors declare no competing interests.

Funding Information: No specific grant was received for this study.

Author Contribution: Sharik Ahmad and Prof. Rafat Parveen jointly contributed to conceptualization, methodology, analysis, and manuscript preparation.

Data Availability: Available from the corresponding author upon reasonable request.

Ethical Approval: Not Applicable.

References

- [1] Rania M. Ghoniem, Abeer D. Algarni, Basel Refky, and Ahmed A. Ewees. Multi-modal evolutionary deep learning model for ovarian cancer diagnosis. *Symmetry*, 13(4):643, 2021. doi:10.3390/sym13040643.
- [2] Matteo Togninalli, Xu Wang, Tim Kucera, Sandesh Shrestha, Philomin Juliana, Suchismita Mondal, Francisco Pinto, Velu Govindan, Leonardo Crespo-Herrera, Julio Huerta-Espino, et al. Multi-modal deep learning improves grain yield prediction in wheat breeding by fusing genomics and phenomics. *Bioinformatics*, 39(6):btad336, 2023. doi:10.1093/bioinformatics/btad336.
- [3] Gouranga Charan, Tawfik Osman, Andrew Hredzak, Ngwe Thawdar, and Ahmed Alkhateeb. Vision-position multimodal beam prediction using real millimeter wave datasets. In *2022 IEEE Wireless Communications and Networking Conference (WCNC)*, pages 2727–2731. IEEE, 2022. doi:10.1109/WCNC51071.2022.9771835.

- [4] Zhong-Hao Ren, Zhu-Hong You, Quan Zou, Chang-Qing Yu, Yan-Fang Ma, Yong-Jian Guan, Hai-Ru You, Xin-Fei Wang, and Jie Pan. Deepmpf: deep learning framework for predicting drug–target interactions based on multi-modal representation with meta-path semantic analysis. *Journal of Translational Medicine*, 21(1):48, 2023. doi:10.1186/s12967-023-03876-3.
- [5] Sandip Jana and Saikat Kumar Shome. Hybrid ensemble based machine learning for smart building fire detection using multi modal sensor data. *Fire Technology*, 59(2):473–496, 2023. doi:10.1007/s10694-022-01347-7.
- [6] Nam D. Nguyen, Jiawei Huang, and Daifeng Wang. A deep manifold-regularized learning model for improving phenotype prediction from multi-modal data. *Nature Computational Science*, 2(1):38–46, 2022. doi:10.1038/s43588-021-00185-x.
- [7] Sharik Ahmad and Rafat Parveen. M3I: Multi-modal machine learning framework integrated with nano-informatics for early prediction of rheumatoid arthritis. *EDPACS*, 71(1):40–62, 2026. doi:10.1080/07366981.2025.2518666.
- [8] Geeta Rani, Meet Ganpatlal Oza, Vijaypal Singh Dhaka, Nitesh Pradhan, Sahil Verma, and Joel J. P. C. Rodrigues. Applying deep learning-based multi-modal for detection of coronavirus. *Multimedia Systems*, 28(4):1251–1262, 2022. doi:10.1007/s00530-021-00824-3.
- [9] Afreen Khan and Swaleha Zubair. An improved multi-modal based machine learning approach for the prognosis of alzheimer’s disease. *Journal of King Saud University - Computer and Information Sciences*, 34(6):2688–2706, 2022. doi:10.1016/j.jksuci.2020.04.004.
- [10] Kh Tohidul Islam, Sudanthi Wijewickrema, and Stephen O’Leary. A deep learning based framework for the registration of three dimensional multi-modal medical images of the head. *Scientific Reports*, 11(1):1860, 2021. doi:10.1038/s41598-021-81044-7.
- [11] Qin Tang, Jing Liang, and Fangqi Zhu. A comparative review on multi-modal sensors fusion based on deep learning. *Signal Processing*, 213:109165, 2023. doi:10.1016/j.sigpro.2023.109165.
- [12] Yeonghun Kang, Hyunsoo Park, Berend Smit, and Jihan Kim. A multi-modal pre-training transformer for universal transfer learning in metal–organic frameworks. *Nature Machine Intelligence*, 5(3):309–318, 2023. doi:10.1038/s42256-023-00628-2.
- [13] Shallu Sharma and Pravat Kumar Mandal. A comprehensive report on machine learning-based early detection of alzheimer’s disease using multi-modal neuroimaging data. *ACM Computing Surveys*, 55(2):1–44, 2022. doi:10.1145/3505244.
- [14] Janani Venugopalan, Li Tong, Hamid Reza Hassanzadeh, and May D. Wang. Multimodal deep learning models for early detection of alzheimer’s disease stage. *Scientific Reports*, 11(1):3254, 2021. doi:10.1038/s41598-020-74399-w.
- [15] Alex Aliper, Roman Kudrin, Daniil Polykovskiy, Petrina Kamyra, Elena Tutubalina, Shan Chen, Feng Ren, and Alex Zhavoronkov. Prediction of clinical trials outcomes based on target choice and clinical trial design with multi-modal artificial intelligence. *Clinical Pharmacology & Therapeutics*, 114(5):972–980, 2023. doi:10.1002/cpt.3008.
- [16] Z. Pang, X. Wang, X. Wang, J. Qi, Z. Zhao, Y. Gao, Y. Yang, and P. Yang. A multi-modal data platform for diagnosis and prediction of alzheimer’s disease using machine learning methods. *Mobile Networks and Applications*, 26(6):2341–2352, 2021. doi:10.1007/s11036-021-01834-1.
- [17] K. Xiong, G. Cao, M. Jin, and B. Ye. A multi-modal deep learning approach for predicting type 2 diabetes complications: early warning system design and implementation. *World Journal of Innovation and Modern Technology*, 7(6):120–132, 2024. doi:10.53469/wjimt.2024.07(06).15.
- [18] Batuhan K. Karaman, Elizabeth C. Mormino, Mert R. Sabuncu, and Alzheimer’s Disease Neuroimaging Initiative. Machine learning based multi-modal prediction of future decline toward alzheimer’s disease: an empirical study. *PLoS ONE*, 17(11):e0277322, 2022. doi:10.1371/journal.pone.0277322.
- [19] Adrienne Kline, Hanyin Wang, Yikuan Li, Saya Dennis, Meghan Hutch, Zhenxing Xu, Fei Wang, Feixiong Cheng, and Yuan Luo. Multimodal machine learning in precision health: A scoping review. *npj Digital Medicine*, 5(1):171, 2022. doi:10.1038/s41746-022-00712-8.
- [20] Deshan Zhou, Zhijian Xu, WenTao Li, Xiaolan Xie, and Shaoliang Peng. Multitdi: drug–target interaction prediction based on multi-modal representation learning to bridge the gap between new chemical entities and known heterogeneous network. *Bioinformatics*, 37(23):4485–4492, 2021. doi:10.1093/bioinformatics/btab473.
- [21] Xiangdong Pei, Ke Zuo, Yuan Li, and Zhengbin Pang. A review of the application of multi-modal deep learning in medicine: bibliometrics and future directions. *International Journal of Computational Intelligence Systems*, 16(1):44, 2023. doi:10.1007/s44196-023-00225-w.
- [22] Shuai Zheng, Zhenfeng Zhu, Zhizhe Liu, Zhenyu Guo, Yang Liu, Yuchen Yang, and Yao Zhao. Multi-modal graph learning for disease prediction. *IEEE Transactions on Medical Imaging*, 41(9):2207–2216, 2022. doi:10.1109/TMI.2022.3159264.
- [23] Binyang Song, Rui Zhou, and Faez Ahmed. Multi-modal machine learning in engineering design: A review and future directions. *Journal of Computing and Information Science in Engineering*, 24(1):010801, 2024. doi:10.1115/1.4063954.
- [24] Md Abdur Rahaman, Jiayu Chen, Zening Fu, Noah Lewis, Armin Iraj, and Vince D. Calhoun. Multi-modal deep learning of functional and structural neuroimaging and genomic data to predict mental illness. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 3267–3272. IEEE, 2021. doi:10.1109/EMBC46164.2021.9630693.
- [25] Kaiwen Tan, Weixian Huang, Xiaofeng Liu, Jinlong Hu, and Shoubin Dong. A multi-modal fusion framework based on multi-task correlation learning for cancer prognosis prediction. *Artificial Intelligence in Medicine*, 126:102260, 2022. doi:10.1016/j.artmed.2022.102260.
- [26] Muyu Wang, Shiyu Fan, Yichen Li, Zhongrang Xie, and Hui Chen. Missing-modality enabled multi-modal fusion architecture for medical data. *Journal of Biomedical Informatics*, 164:104796, 2025. doi:10.1016/j.jbi.2025.104796.

- [27] Luhui Cai, Weiming Zeng, Hongyu Chen, Hua Zhang, Yueyang Li, Yu Feng, Hongjie Yan, Lingbin Bian, Wai Ting Siok, and Nizhuan Wang. Mm-gtunets: Unified multi-modal graph deep learning for brain disorders prediction. *IEEE Transactions on Medical Imaging*, 44(9):3705–3716, 2025. doi:[10.1109/TMI.2025.3556420](https://doi.org/10.1109/TMI.2025.3556420).
- [28] Lei Wang and Yigang He. M2stan: Multi-modal multi-task spatiotemporal attention network for multi-location ultra-short-term wind power multi-step predictions. *Applied Energy*, 324:119672, 2022. doi:[10.1016/j.apenergy.2022.119672](https://doi.org/10.1016/j.apenergy.2022.119672).
- [29] Vasileios Skaramagkas, Anastasia Pentari, Zinovia Kefalopoulou, and Manolis Tsiknakis. Multi-modal deep learning diagnosis of parkinson's disease—a systematic review. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:2399–2423, 2023. doi:[10.1109/TNSRE.2023.3277749](https://doi.org/10.1109/TNSRE.2023.3277749).



Sharik Ahmad received his Ph.D. in Computer Science and is currently pursuing a second Ph.D. at Jamia Millia Islamia, New Delhi, India. His research interests include Artificial Intelligence, Machine Learning, Multimodal Learning, Healthcare Analytics, Large Language Models, and Data Science. He has published several research papers in SCI, Scopus, and Web of Science indexed journals and conferences.



Rafat Parveen is a Professor in the Department of Computer Science at Jamia Millia Islamia, New Delhi, India. Her research focuses on Bioinformatics, Machine Learning, Cloud Security, and Computational Systems Biology. She has published extensively in reputed journals and conferences and has supervised numerous Ph.D. scholars.