JCBI
Journal of Computing & Biomedical Informatics
ISSN: 2710 - 1606
Volume 11 Issue 01
2026

Research Article
https://doi.org/10.56979/1101/2026/1257

An ANOVA-SMOTE-Based Framework for Multi-Class Autoimmune Disease Classification

Khadija Kanwal[1], Syeda Noor-ul-ain Bukhari[1], Ayousha Tahreem[1], Sumayya Suleman[1], Fariha Zulfiqar[1], and Aiza Shabir[1*]

[1]Institute of Computer Science and Information Technology, The Women University Multan, Pakistan.
[*]Corresponding Author: Aiza Shabir. Email: aiza.6322@wum.edu.pk

Received: January 07, 2026 **Accepted:** April 20, 2026

Abstract: Diagnosis of autoimmune disorders is prone to significant difficulties as the clinical picture is overlapped with complicated pathophysiology. The proposed study is the multi-class classification of autoimmune diseases based on a machine learning framework and taking into account four disease categories, namely, Graves' Disease, Sjögren Syndrome, Rheumatoid Arthritis, and Systemic Lupus Erythematosus as well as a healthy control group, which leads to a five-class classification problem. The framework takes advantage of an integrated dataset, which includes demographic, hematological, biochemical, as well as immunological characteristics. K-Nearest Neighbors (KNN), Linear Support Vector Classifier (SVC), Decision Tree, Naive Bayes, Logistic Regression, and Ridge Classifier are six classification algorithms that were systematically tested under the same conditions of experiment. In order to improve the performance of the models and solve the problem of the class imbalance, an ANOVA-based feature selection strategy was utilized to be combined with the Synthetic Minority over Sampling Technique (SMOTE). The findings of experimental studies confirm that nonlinear classifiers, especially, Decision Tree and KNN, are much better than the linear models in terms of F1-score, recall, and precision. The added feature of the framework is the incorporation of a healthy control group, which enhances the capacity of the framework to differentiate between a diseased and a non-diseased person and increase the clinical applicability. In general, the suggested ANOVA-SMOTE-based framework can be considered as a powerful and scalable method of non-invasive autoimmune diseases classification, which can be implemented in the real clinical diagnostics and decision support.

Keywords: Autoimmune Diseases; Machine Learning; Classification; SMOTE; ANOVA Feature Selection; Medical Diagnosis; Decision Tree; K-Nearest Neighbors; Graves' Disease; Sjögren Syndrome; Rheumatoid Arthritis; Systemic Lupus Erythematosus

1. Introduction

Autoimmune diseases occur when the immune system erroneously recognizes healthy cells, tissues, and body organs as foreign and causes an improper immune reaction towards the body parts. Such conditions are costly to the world in terms of health, as they impact millions of people globally [1]. The pathophysiology is the presence of chronic inflammation and progressive tissue damage, as the result of dysregulated immunogenicity. Autoimmune diseases may attack almost all organ systems, such as the joint system, the skin, nervous system, heart system, kidneys, gastro-intestinal system, connective tissues and glands. It is very complicated with interactions between immune cells, hereditary disposition, and the environment [2]. At least 15 diseases are known to be directly caused by autoimmune mechanisms and many other diseases have autoimmune constituents [3]. Among the most well-known ones are rheumatoid arthritis (RA), systemic lupus erythematosus (SLE), type 1 diabetes mellitus (T1DM) and multiple sclerosis (MS). Although much research has been done, the exact etiologies of most of the autoimmune diseases

have not been established yet. The existing knowledge indicates that the autoimmunity is a consequence of the impaired immune tolerance, the disturbed recognition of antigens and the tissue particular vulnerability [4]. Furthermore, autoimmune endocrine diseases can also be comorbid with the autoimmune polyglandular syndromes or be seen together with non-endocrine autoimmune diseases [5]. Autoimmune diseases are on the rise especially in developed nations, where they are known to afflict about 5-10 percent of the population and have high morbidity and mortality as well as health expenditure rates equivalent to the cardiovascular and cancer diseases [6]. These disorders may arise at any age and in most cases show a clustering within a family with a greater concordance rate of identical twins [7]. Interestingly, the proportion of cases in women is close to 79% which indicates a high gender inequality which is a result of genetic, hormonal, and environmental factors [3]. Even with the current medical research progress, the early and faithful diagnosis of the autoimmune diseases is still difficult because of the similarity in clinical manifestations, the heterogeneity of disease course, and the absence of the highly specific biomarkers. Such issues indicate the necessity of highly developed and data-driven diagnostic methods. Our topic in this paper is on four clinically important autoimmune diseases Graves' Disease (GD), Sjögren Syndrome, Rheumatoid Arthritis (RA), and Systemic Lupus Erythematosus (SLE). These diseases were chosen according to their prevalence rate, clinical significance and access to validated diagnostic indicators. SLE is an autoimmune disorder that is persistent and multi systemic with the production of auto antibodies like antinuclear antibodies (ANA) and anti- double-stranded DNA (anti-dsDNA) that attack various organs. Its various clinical presentations such as skin rashes, pains, and tiredness, and inflammation of body organs usually make it difficult to diagnose because of its resemblance with other diseases [8]. Likewise, RA is an autoimmune disease characterized by the progressive one that affects mostly the joints resulting in destruction of cartilage and bones. Whereas biomarkers like rheumatoid factor (RF) and anti-cyclic citrullinated peptide (anti-CCP) antibodies will help in the diagnosis, the lack of absolute markers tends to delay early diagnosis [912]. RA should be treated with a pharmacological and non-pharmacological approach to prevent disability and enhance the quality of life [13]. Sjögren's Syndrome is marked by failures of exocrine glands and this leads to the mouth, eyes as well as other mucosal surfaces drying up. It is traditionally linked with increased immunoglobulin and auto-antibodies including anti-Ro/SSA and anti-La/SSB [1416]. Recent research indicates that anti-Ro52 antibodies may interfere with immune regulation processes, which leads to the presence of chronic inflammation [17]. The condition might be isolated or be accompanied with other autoimmune disorders like RA and SLE. The Graves' Disease (GD), which is one of the major causes of hyperthyroidism, is most common among women of reproductive age [18]. It presents itself in symptoms of weight loss, heat intolerance, tachycardia and the enlargement of the thyroid gland and some patients getting Graves' ophthalmopathy. Diagnosis is usually done through identification of thyroid-stimulating hormone receptor antibodies (TRAb) and increased or decreased levels of thyroid hormone usually assisted by imaging, including thyroid scintigraphy. Despite the availability of anti-thyroid drugs, radiological iodine therapy and thyroidectomy as treatment, there are still high chances of recurring the disease, and osteoporosis and cardiovascular diseases can be experienced in the long term.

In order to solve the diagnostic challenges posed by the autoimmune diseases, this study presents the problem in multi-class classification in the form of five classes, where 4 of the classes constitute autoimmune diseases (GD, Sjögrens Syndrome, RA and SLE) with the remaining 1 class as a healthy control group. The healthy class is included, and it is important to note that it allows the model to separate diseased and non-diseased people effectively and, thus, increase its clinical applicability. We are offering a unified machine learning system, which combines demographic information, hematological metrics, biochemical and immunological indicators in one pipeline. ANOVA feature selection is utilized to select the most discriminative features to enhance the model performance, and Synthetic Minority over Sampling Technique (SMOTE) is employed to overcome the problem of class imbalance. Multiple machine learning techniques were tested with cross-validation to increase model stability and ability to generalize. It is the aim of this framework to increase the diagnostic accuracy and model interpretability to help with clinical decision making and the ability for earlier diagnosis of autoimmune diseases leading to a better patient outcome.

Table 1. Overview of autoimmune diseases research

Disease	Affected System	Key Autoantibodies	Common Symptoms
Rheumatoid Arthritis	Synovial joints	RF, Anti-CCP	Joint swelling, stiffness, pain, deformities
Systemic Lupus Erythematosus	Multi-systemic(joints, skin, kidneys, CNS)	Anti-dsDNA, ANA	Fatigue, joint pain, rashes, organ inflammation
Sjögren's Syndrome	Exocrine Glands(salivary, lacrimal)	SSA(Ro),SSB (La)	Dry mouth, dry eyes, joint pain, fatigue
Graves' Disease	Thyroid gland	TSH receptor antibodies	Weight loss, heat intolerance ,anxiety, goiter

Here, we demonstrate the first multi-class classification approach to diagnosing GD, Sjögren syndrome, RA, and SLE through the application of a machine learning pipeline with features including demographic information, hematologic measurements, and autoantibody markers. With ANOVA for feature selection and SMOTE for imbalance, we perform systematic cross validation of both traditional ML techniques and deep neural networks. The objective is to increase the interpretability and diagnostic precision of the obtained model and to establish it as a foundation for clinical decision support systems.

2. Literature Review

Diseases that belong to the group of autoimmune diseases are challenging for diagnoses, due to the similar symptoms and complex etiopathogenesis. While the traditional laboratory test including blood count and autoantibodies detection (such as ANA, anti-dsDNA, RF, anti-CCP) is required, this method often cannot sufficiently determine the classification for early diagnosis.

2.1. ML-Based Diagnosis of Autoimmune Diseases

Based on the systematic literature review over 169 research papers [19], Random Forest and Support Vector Machine (SVM) were extensively utilized to identify autoimmune disease states. However, less than 10% research applied strict cross-validation procedures or incorporated multi-source data, revealing methodological deficiency in the area. More recent research studies focused more on identifying significant biomarkers with machine learning approaches, so that they can promote diagnoses in autoimmune diseases. For instance, the research of Zhang et al. Utilized ML classifiers for identifying blood biomarkers with strong predicting capability, and achieve good performance. The paper demonstrated that ML approach is capable to conduct early risk stratification by identifying significant blood biomarkers for diagnosing ILD, which is an acquired pulmonary condition from Rheumatoid Arthritis [20].

The ensemble methods are also performed well for predictive modeling in autoimmune diseases in relation to thyroid. The ensemble models (such as Random Forest and Gradient Boosting) showed its ability to predict Hashimoto's Thyroiditis accurately based on both clinical and laboratory measures, with comprehensive evaluation from en et al. [21]. Regarding discriminate of RA and SLE on genomic studies, models based on genome-wide SNP data perform better than logistic regression on distinguishing RA from SLE. For example, Random Forests classified RA from SLE with a resulting AUC of 0.98. As for feature importance, the key variables were the HLA-DQA1, HLA-DQB1 and HLA-DRB1 allele.

The deep learning methods have also been proposed for the autoimmune diseases in addition to traditional machine learning approaches [32]. A CNN-LSTM hybrid model combined with attention fusion layer was proposed for classifying SLE related epitope sequences. It achieves ROC-AUC of 0.9506. Both of the architectures (CNN and LSTM) were integrated to combine the power of the CNN-based spatial representations and LSTM-based time dependencies. Furthermore, attention mechanism enhances the classification prediction performance by focusing on most discriminative epitope features. Interpretability of such model is useful to pinpoint subtle immunologic features, thus improves identification and understanding of autoimmune diseases such as SLE and RA which have complex clinical presentations

[22]. Such hybrid models are the possible way for the identification accuracy of autoimmune diseases on high dimensional biomedical data.

2.2. Modeling Strategies for Autoimmune Disease Analysis

Several studies created ML based diagnosis methods for Systemic Lupus Erythematosus (SLE) on standard clinical and serologic measures. For instance, with external testing cohort, a Random Forest classifier on 8 lab testing (e.g. Uric acid, Thrombin time) resulted in an AUC of 0.83. Similarly, gradient boosting algorithms (e.g. XG Boost) performed well with nearly 90% balanced accuracy [19]. All these results indicate that the ensemble approach can better handle the heterogeneous biomedical data to predict autoimmune diseases accurately. Cost-sensitive models and ANNs have also demonstrated promise in RA. One remarkable model using age and anti-Carp antibodies achieved 90.6% accuracy which surpassed threshold-based models. Deep learning on ultrasound imaging has shown comparable result to that of expert clinicians (AUC up to 0.95).

ML paradigms for the detection of autoimmune diseases have been comprehensively compared in several researches. In one of such comparison, the performance of various classification algorithms such as SVM, Random Forest and Decision Trees to detect autoimmune diseases were compared [31]. The comparative studies provide insightful comparison for model selection in diagnosis tasks. The results revealed tree-based algorithms and ensemble models outperform linear classifiers in terms of accuracy and generalization, which confirms the efficiency of nonlinear model to handle complex interrelationships between features in autoimmune disease data [23].

The application of ML and DL algorithms to diagnosing and discriminating autoimmune diseases has been a major area of research in recent years. These techniques have been utilized on various types of data including imaging modalities, hematologic measurements, hormone levels, clinical records, and genetic information. A variety of models have been proposed for discriminating autoimmune diseases such as rheumatoid arthritis, systemic lupus erythematosus, Sjögren syndrome and Graves' disease, ranging from traditional ML classifiers such as support vector machines and logistic regression to more sophisticated deep learning models such as LSTMs and CNNs. A comprehensive comparison of some relevant recent studies can be found in Table 2 including the proposed methods, data sources, performance indices, and key discriminative variables reported.

Apart from emphasizing the potential for data-driven approaches for the diagnosis of autoimmune diseases, this comparison also provides a useful snapshot for the existing limitations which includes multi-class classification, feature interpretability, and generalizability to new patient populations.

Table 2. Comparative Summary of Existing Studies on Autoimmune Disease Diagnosis Using ML and DL

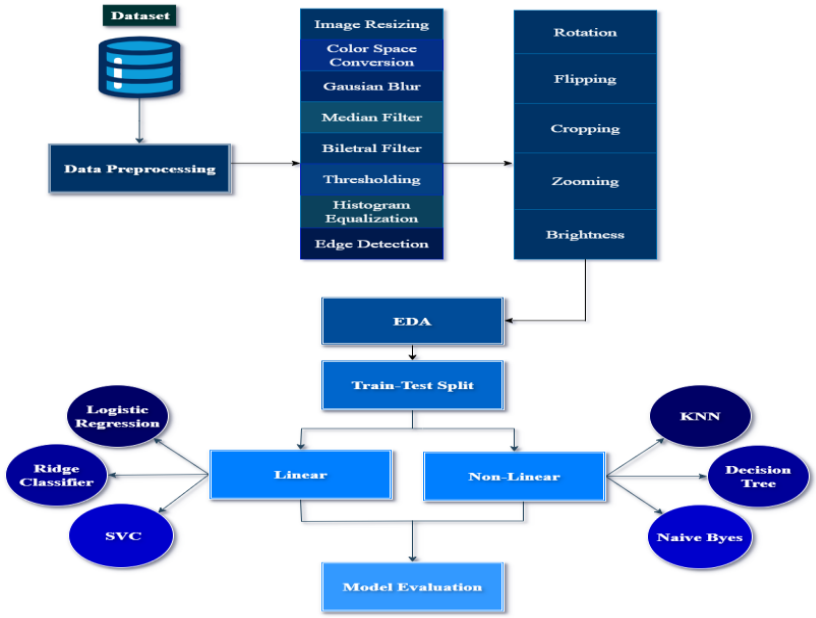
Author (Year)	Model/ Technique	Target Disease(s)	Data Type/ Source	Accuracy	Limitations
Liu et al.(2021)	Random Forest	SLE	Clinical and Immunological	89%	Focused only on SLE (limited scope).Relied on clinical & immunological data only
Chen et al.(2020)	Deep Neural Network	RA	Symptom +Bio marker	92%	Focused only on RA (narrow scope) Small dataset (risk of overfitting)
Wu et al.(2022)	XG Boost	Hashimoto's Thyroiditis	CBC + Autoantibodies	91%	Disease-specific (only HT, limited generalization) Lacked explainability for

					clinical use (narrow features)
Zhang et al. (2021)	Logistic Regression+ KNN	RA, Graves', Sjögren	EHR, Structured Data	88%	Limited to structured EHR data (missed unstructured notes/imaging) No external validation for generalizability
Singh et al.(2019)	SVM + Decision Tree	RA, Hashimoto	Auto antibodies, CBC	85%	Lower accuracy (85%) compared to newer models Small dataset, increasing risk of overfitting
Gupta et al.(2023)	CNN + Transfer Learning	Autoimmune Skin Diseases	Dermoscopic images	94%	Focused only on skin-related autoimmune diseases (narrow scope) Lack of integration with serological or genetic features
Bai et al.(2022)	Light GBM	SLE,RA	Genomic & Immunologic	90%	Focused only on SLE and RA (limited disease coverage) Possible overfitting due to high-dimensional genomic features
Al-Marzouqi et al.(2020)	Naïve Bayes	Sjögren Syndrome	Questionnaire+ CBC	84%	Relied on questionnaire data → subject to bias and variability No external or multi-center validation
Lee et al.(2021)	Voting Classifier	SLE, RA, Hashimoto, Sjögren, Graves	CBC, hormone, and Auto immune markers	Up to 93%	Lack of longitudinal or temporal data analysis Did not evaluate real-world clinical deployment

Ahmed et al.(2022)	BiLSTM	SLE	RNA-seq Gene Expression	95%	Small and domain-specific RNA-seq dataset Risk of overfitting without extensive cross-validation
Fatima et al. (2023)	CNN+LSTM	Multi-class Autoimmune	CBC + Antibody Dataset	96%	Dataset size may be insufficient for deep hybrid CNN + LSTM architecture Model interpretability not addressed, hindering clinical adoption
Rahman et al.(2022)	Decision Tree	Graves' Disease	Hospital records	89%	Single-center hospital records, limiting generalizability Lack of integration of diverse biomarkers (focus only on records)
Khan et al.(2021)	XG Boost	RA, Hashimoto, Graves	Clinical bloodwork	91%	Reliance only on clinical blood work, missing genetic or imaging data No longitudinal or temporal data considered

3. Materials and Methods

The present study endeavors to employ supervised machine learning techniques to classify autoimmune disorders based on serological and clinical features. The structure of the proposed method is organized into 6 steps starting with data collection and preprocessing and then followed by Exploratory-Data-



Analysis (EDA), feature engineering and selection, model development and training, and evaluation. These steps were designed to ensure data validity, increase the reliability of the results and align them with clinical relevance.

Figure 1. Schematic Representation of the Methodological Pipeline for Autoimmune Disease Prediction.

3.1. Dataset Insights

Data was collected and organized from 12,500 patients comprising a diverse spectrum of healthy individuals and patients diagnosed with various autoimmune conditions, based on their clinical and serological features. These were then prepared with the intent to aid classification and diagnosis of these conditions using computational learning models.

The dataset incorporates three main types of attributes; the first is demographic information that consists of an individual's age and gender, for instance. The second attribute category comprises Hematological information that is extracted from Complete Blood Count (CBC) tests and this comprises data such as Erythrocyte count (RBC), Leukocyte count (WBC), Average Corpuscular Hemoglobin (MCH), Average Red blood cell volume, HB level, Hematocrit, Mean Corpuscular Hemoglobin concentration (MCHC) and platelet levels.

The third type comprises key serological attributes that is used to help differentiate between autoimmune reactions and this comprises rheumatoid factor (RF), antinuclear antibodies, antibodies against cyclic citrullinated peptides, anti-double-stranded DNA among others. The corresponding classification of these attributes is a confirmed clinical diagnosis allowing for multi-class classification between a healthy individual and various autoimmune diseases such as thyroid associated autoimmune disease, system Lupus, Sjögren's syndrome, and autoimmune arthritis in the joints.

Table 3. Dataset Feature Description

Features	Measure	Description	Count
		Input	
Age	Y (Years)	Age of the patient	12,500
Gender	N/A	Gender of the patient	12,500
BMI	BMI, kg per square meter	Body Mass Index	12,500
Glucose	Concentration in milligrams per deciliter(mg)	Fasting glucose level	12,500
Insulin	Units per milliliter	Insulin hormone level	12,500
Testosterone	Nano grams per deciliter	Testosterone level	12,500
HbA1c	(Percentage)%	Glycated hemoglobin(long-term glucose control)	12,500
WBC Count	($\times 10^3$ per μL of blood)	White blood cell count	12,500
RBC Count	($\times 10^6$ per μL)	Red blood cell count	12,500
HB Level	(grams per deciliter)	Hemoglobin level	12,500
Hct (%)	(Percentage)%	Hematocrit(volume % of red blood cells)	12,500
MCV (fL)	(fem to liters, fL)	Mean Corpuscular Volume	12,500
MCH (pg)	(picograms, pg)	Mean Corpuscular Hemoglobin	12,500

MCHC (gdL)	(grams/dL)	Mean Corpuscular Hemoglobin Concentration	12,500
Platelets Counts	(×10³/μL of blood)	Platelet count	12,500
ANA	-	Anti-Nuclear Antibodies	12,500
Anti-dsDNA	-	Anti-double stranded DNA antibodies	12,500
RF	IU per mL	Rheumatoid Factor	12,500
Anti-CCP	U per mL	Anti-cyclic citrullinated peptide	12,500

.

Table 4. Output Classes

Class Label	Class Name	Description	Sample Count
2	Rheumatoid Arthritis (RA)	Chronic autoimmune joint disease	2,192
1	Systemic Lupus Erythematosus (SLE)	Multisystem autoimmune disease	2,490
3	Graves' Disease (GD)	Hyperthyroidism based Autoimmune	130
5	Hashimoto	Hashimoto's Thyroiditis	2,200
4	Sjögren's Syndrome	Autoimmune exocrine gland disorder	2,390
	Healthy Control	No autoimmune disease	3,098

3.2. Data Preprocessing

Machine learning algorithms were used on the dataset after a set of preprocessing operations were applied to assure data consistency and quality. These operations were used to fix problems like missing data, mixed data types of categories, imbalances in features, and duplicated entries for a more reliable and reliable model.

Dataset Sanitization: The data set was inspected for null or missing data and either removed if it was unrecoverable or substituted using methods of the mean or mode. Irrelevant columns such as patient ID were discarded because they could not be learned from.

Encoding categories: Categories like sex (male/female) were converted appropriately. Target variables (different autoimmune disease classifications and controls) were encoded using label encoding.

Scaling features: Using the "Standard Scaler," the dataset was standardized. All numerical attributes were given the same relative importance when learning; this prevented disparities from the range or magnitude of each feature from interfering with the effectiveness of the model. By using the Standard Scaler, features are adjusted to a mean of 0 and a standard deviation of 1 and a higher classification prediction accuracy resulted [24].

Dimensionality reduction: The Variance Threshold function was utilized to remove low variance data which was assumed to provide minimal additional information for classification. A threshold value of 0.01 was implemented to only keep features with sufficient variation across the samples.

Final feature set: At this point, the demographic data features such as the age and sex of the patient are available, along with the blood parameters, such as count for total erythrocyte, whites and red blood cell, hb, Packed cell volume/PCV, Average red blood cell volume, Average hemoglobin per cell, Average corpuscular hemoglobin concentration, and Thrombocyte count, and autoimmune markers, ANA, dsDNA-specific autoantibodies, RF, Autoantibodies targeting cyclic citrullinated peptides. The specific clinical outcome was recorded for each patient. This pipeline increased computational speed, minimized data noise and redundant data, and enhanced the overall ability of the classification models.

While Table 3 shows the main clinical, biochemical, hematological and immunological characteristics utilized in this study, the full dataset also encompasses an augmented list of disease-specific immunological markers as well as various hematological sub-parameters. Specifically, the list includes:

Anti-BP180, Anti-BP230, C1 inhibitor, anti-tissue transglutaminase, MBL level, ASCA, Anti-Tg, Anti-CBir1 as well as different parameters relating to hematological profile including eosinophils and platelet parameters.

During the process of feature selection, the extended feature list was considered to improve the discriminative power with respect to each disease. Thus, the ANOVA feature ranking shown in Table 4 was derived from the entire list of features, which included the baseline characteristics in Table 3 in addition to the immunological parameters. The parameters in Table 3 have now been re-edited to show the baseline characteristics, and Table 4 demonstrates the set of the most significant features identified from the full feature list.

Table 5. Top 10 Features by ANOVA F-value

Rank	Features	F-Value	P-Value
1	Anti_BP180	71.95	3.86×10^{-74}
2	Anti_BP230	71.95	3.86×10^{-74}
3	C1_inhibitor	53.72	3.53×10^{-55}
4	Anti-tissue transglutaminase	52.68	4.32×10^{-54}
5	MBL Level	41.37	2.88×10^{-42}
6	ASCA	35.22	8.02×10^{-36}
7	Anti-Tg	35.10	1.07×10^{-35}
8	PLT Count	33.31	8.07×10^{-34}
9	Eosinophils	32.28	9.57×10^{-33}
10	Anti_CBir1	30.18	1.52×10^{-30}

3.3. Model Development and Training

To efficiently diagnose the five autoimmune diseases, this study proposed a ML pipeline that integrates clinical features and serological data. Some fundamental operations in constructing and training the ML model are given below:

3.3.1. Data Splitting

For building our model the available data was divided into 80:20 as training and test set. Considering the imbalance that existed in the original data where only 130 cases of Graves' Disease (GD) were available as against 2490 cases of Systemic Lupus Erythematosus (SLE) and 2390 cases of Sjögren Syndrome (SS) a stratified sampling was used in order to have a proportion of each class including the under-represented class of GD in the training as well as test set to avoid evaluation bias.

To overcome class imbalance a SMOTE-based augmentation method was used for the training set only which creates artificial samples based on interpolating values between minority samples and their neighboring samples and not just copying the existing samples, in order to balance the number of classes in the training set without any effect on test set and to avoid data leakage.

After augmenting the training set it was used for training the model and tested on the original unseen test set. As per experimental results presented here the proposed framework is effective in distinguishing between healthy controls and different autoimmune classes, the accuracy, precision, recall and f1-score values are relatively stable for all classes.

Although the class imbalance presented a problem which mostly affects the GD class, by performing stratification in the training-testing split and data augmentation on training data, it significantly helped in better generalization of the model. Small variations were seen in the values across different classes which is natural in imbalanced multi-class classification problems.

3.3.2. Training of Models

The aim of this study was to diagnose the autoimmune diseases with high accuracy, it explored the use of multiple ML models, proposing a unified ML classification pipeline by leveraging the 15 highly discriminative features from ANOVA F-test, the features were then heavily pre-processed by Standard

Scaler in normalization, and SMOTE to address extreme class imbalances to ensure equal class representation for the training data set.

Six classification algorithms viz Decision Tree, Logistic Regression, Ridge Classifier, Linear Support Vector Classifier (SVC), K-Nearest Neighbors (KNN) and Naive Bayes classifier are compared in the same experimental setup. It shows that nonlinear models with structure awareness are more adept at capturing the complex interplay of clinical features associated with autoimmune etiology.

These rigorous processes result in a robust starting point for the incorporation of explainable AI into autoimmune diagnosis, also establishing the predictive power of tree-based and distance-based learners.

3.3.3. Mathematical Framework for Linear Classifiers in Autoimmune Disease Prediction

Three linear models, namely Logistic Regression, Ridge Classifier, and Linear Support Vector Classifier (SVC) are incorporated to serve as baselines for classification accuracy because of their interpretability and mathematical tractability. Logistic Regression leverages the sigmoid function to model the log-odds as a linear combination of the input features and then applies the inverse link function to obtain the class probabilities [27]. The Ridge Classifier improves upon logistic regression by introducing L2 regularization to shrink coefficient estimates and prevent overfitting. Linear SVC identifies the optimal hyperplane in the feature space that maximizes the margin between classes [29]. It assumes linear separability in data. Although both these methods are robust in many classification scenarios, the underlying assumption of a linear decision boundary may prove to be inadequate to describe the complicated nonlinear association between disease etiology and the various clinical measurements, such as ANA, hematocrit, or ESR, although it does support probability interpretations on certain features.

The Ridge Classifier is used to prevent overfitting and increase the stability of the coefficients because it is well known that high-dimensional data may have issues with multi-collinearity. L2 regularization term is added in the model which shrinks excessive weights, providing a smoother decision boundary [28]. The performance of Ridge model shows slight improvement even though it only marginally increases the predictive accuracy and is yet incapable of capturing latent feature hierarchies in the autoimmune feature profiles [33]. The Linear Support Vector Classifier(SVC) maximizes the distance between class labels [29]. Similarly, like logistic regression, it cannot model the inherent nonlinearities in the relationship of the variables with the disease, nor does it perform well in cases with substantial overlap and high variance between classes, common in the context of autoimmune diseases. It demonstrates that classical linear models may fail to provide good predictions for this set of data and highlights the need of non-linear models. Mathematical formulation for logistic regression is given in equation 1:

$$\hat{y} = f(\mathbf{x}) = \sigma(\mathbf{w}^{\top}\mathbf{x} + b), \quad \text{with} \quad \min_{\mathbf{w},b}[\mathcal{L}(\mathbf{w}, b) + \lambda\|\mathbf{w}\|_2^2] \tag{1}$$

Where $\sigma(z) = \frac{1}{1+e^{-z}}$ and $\mathcal{L}$ is the loss function

3.3.4. Mathematical Framework for Nonlinear Classifiers in Autoimmune Disease Prediction:

To tackle the complexity of nonlinear relationships, three non-parametric and probabilistic classifiers namely KNN, Naive Bayes and Decision Tree were incorporated with the above three linear models. Each model uses different principles and have different aptitude in finding out individual patterns in the patients: K-Nearest Neighbors (KNN) is a metric based approach. It finds the k nearest training examples based on the Euclidean distance and predicts the label of unknown instance. For autoimmune diseases, such an approach is meaningful because local similarity in the feature space means a similar disease profile (based on the variables), and such local patterns in clinical symptoms could imply therapeutically implications in equations 2 and 3.

$$d(s,t) = \sqrt{\sum_{x=1}^{n}(s_x - t_x)^2} \tag{2}$$

$$\hat{y} = \text{mode}\left(y_i \mid t_i \in N_k(s) \right) \tag{3}$$

Decision Tree classifier recursively divides the training set based on a chosen feature that minimizes the impurity of the resulting child nodes, until a stopping criterion is met. The division is based on metrics like information gain or Gini impurity in equation 4 and 5. In a similar fashion physicians approach differential

diagnosis of autoimmune diseases. One can construct rules such as the thresholds for ESR, ANA or hemoglobin levels, that lead to the classification of particular diseases.

$$\text{Gini}(t)=1-\sum_{i=1}^{C} p_i^2 \quad (4)$$

$$IG(t,a) = H(t) - \sum_{v \in \text{Values}(a)} \frac{|t_v|}{|t|} H(t_v) \quad (5)$$

Naive Bayes classifier based on Bayes Theorem calculates the posterior probability for each class of autoimmune disease given a particular patient, assuming all features including hematological and serological parameters are conditional independent [30]. It integrates the conditional probability of observing the features with prior probabilities of each class, to deliver a robust multi-class classifier, it allows reasonable interpretation and effective handling of high dimensional data. Its applicability extends to modeling the probabilities of observed continuous laboratory parameters with a Gaussian probability density function. Naive Bayes is often known for good accuracy in high-dimensional biological problems, especially when stability and minimal training set size are important factors [30]. Equations are as given below:

$$P(y \mid \mathbf{x}) \propto P(y) \prod_{i=1}^{n} P(x_i \mid y) \quad (6)$$

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2} \right) \quad (7)$$

3.4. Model Evaluation

Different metrics were used to evaluate the performance of the system as it shows the classification accuracy, recall, precision and F1-score of the conventional ML models. Using confusion matrices, and cross validation ensures that all the classes are thoroughly evaluated.

3.4.1. Evaluation Metrics

The metrics used are to measure the accuracy of ML classifiers predictions. They clearly determine the precision, consistency and the generalization ability of the system; and help in a better analysis of the models in classifying various autoimmune diseases and normal subjects in the dataset. The evaluation metrics are calculated for all ML classifier:

Accuracy is defined as the percentage of the total predictions that the classifier correctly identified. The accuracy is expressed by Equation 8.

$$\text{Accuracy} = \frac{\text{TN (True Negative)} + \text{TP (True Positive)}}{\text{TN(True Negative)} + \text{TP(True Positive)} + \text{FN(False Negative)} + \text{FP(False Positive)}} \quad (8)$$

Precision reflects the model's ability to correctly predict positive outcomes in consideration to the class imbalance, where precision is defined by Equation 9:

$$\text{Precision} = \frac{\text{TP (True Positive)}}{\text{FP (false Positive)} + \text{TP (True Positive)}} \quad (9)$$

Recall reflects a model's ability to successfully find all relevant data instances, Recall is defined as follows Equation 10:

$$\text{Recall} = \frac{\text{TP(True Positive)}}{\text{FN(False Negative)} + \text{TP(True Positive)}} \quad (10)$$

F1-score, is the balanced statistic, which considers both false positives and false negatives; it is a combination of both precision and recall values (harmonic mean). The F1-Score is defined as follows Equation 11:

$$\text{F1} - \text{score} = 2 \times \frac{\text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \quad (11)$$

4. Experimental Outcomes

This section presented an evaluation of the accuracy, precision, recall and F1-score of the developed ML models to categorize the autoimmune diseases based on immunological, hematological, biochemical and demographic factors. The performance of each model was assessed against every individual illness class as well as overall for each model. These metrics help to understand the ability of each classifier in distinguishing between different diseases, and healthy controls.

4.1. Dataset Description and Demographics

The dataset consisted of 12,500 records, including the healthy controls and the patients, 2,500 for each autoimmune disease (RA: Rheumatoid arthritis, SLE: Systemic Lupus Erythematosus, SS: Sjögren Syndrome, GD: Graves' Disease). For each record, there were 79 variables encompassing immunological, biochemical, hematological and demographic information (age, gender) that are presented. The gender based trends in the dataset are obvious. For SLE, Sjögren Syndrome and RA groups, it is dominated by female subjects, this is also in line with their relative prevalence in actual life; also in Graves' disease, female subjects are more frequent than male subjects but not to the same ratio; there is the majority gender balance that allows a fair model training as for healthy group. The majority of subjects within the data set are within the ages range between 30-60 which fits with the average onset of most autoimmune diseases.

The dataset used in this study comprises a total of 79 variables: demographic, clinical, biochemical, hematological, and immunological. Table 3 highlights only a core subset of the included variables due to length; these include the key clinical measurements (age, gender, BMI), routine biochemical markers (glucose, insulin, testosterone, HbA1c), routine hematological parameters (CBC values), and main immunological markers (ANA, anti-dsDNA, RF, anti-CCP). The full dataset includes additional specialized immunological (disease specific) markers as well as more refined hematological (CBC sub-parts) parameters including: Anti-BP180, Anti-BP230, C1 inhibitor, anti-tissue transglutaminase, MBL level, ASCA, Anti-Tg, Anti-CBir1, platelet sub-values, and eosinophil count. These variables are all included in the feature selection and model construction to allow disease specific discrimination. As a result, the ANOVA-based ranking (Table 4) used all 79 variables while Table 3 only focuses on the smaller set described above for descriptive reasons.

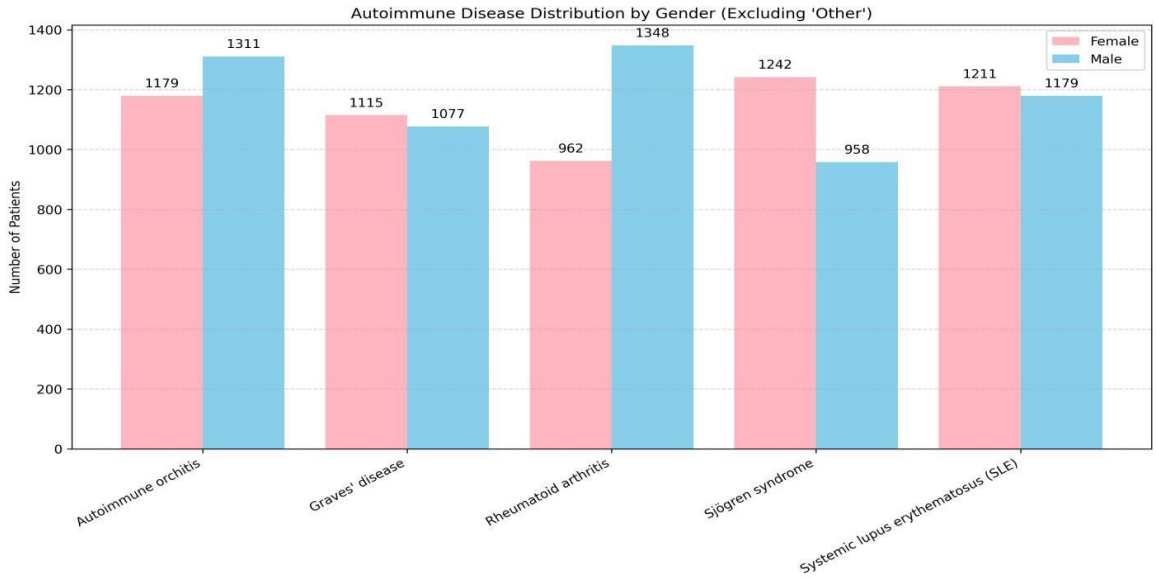


Figure 2. Autoimmune Disease Distribution by Gender.

4.2. Model Evaluation and Performance Metrics

This part of the paper describes the findings of the analysis of various machine learning models that were used to predict the balanced autoimmune disease dataset. This was done with the primary aim of performing multi-class classification over five categories which are four autoimmune diseases namely; Rheumatoid Arthritis (RA), Systemic Lupus Erythematosus (SLE), Sjögren's Syndrome (SS) and Graves' Disease (GD) and a healthy control group using a wide range of immunological, hematological, biochemical and demographical characteristics.

In order to evaluate the performance of models, standard classification measures were used, such as Accuracy, Precision, Recall and F1-score. The metrics are of special significance in medical diagnostics where it is not only necessary to gain high accuracy, but also to be able to guarantee the high reliability of diagnosing specific disease cases and reducing the number of false negative and false positive cases.

The machine learning algorithms were tested as follows:

Logistic Regression (LR): The linear model is used to estimate the likelihood of membership to a specific class based on the use of a logistic function. K-Nearest Neighbors (KNN): This is a non-linear algorithm and a distance-based algorithm which classifies samples according to the majority label of the nearest samples.

Linear Support Vector Classifier (Linear SVC): This is a margin-based linear classifier that classifies the classes with the help of a linear decision boundary.

Decision Tree (DT): This is a non-linear model that recursively divides the feature space according to the decision rules.

Naive Bayes (NB): This is a probabilistic classifier that uses the Bayes theorem with the assumption that different features are conditionally independent of each other.

Ridge Classifier (RC): This is a type of linear classifier which uses L2 regularization to minimize overfitting.

The experimental findings reveal that non-linear models especially Decision Tree and KNN performed better than linear methods. The models achieved an F1-score of more than 0.96 in most classes, which suggests that they have a strong ability to learn complex, and non-linear association between clinical features. However, when it comes to their ability to differentiate between clinically similar conditions like Sjögren Syndrome and SLE, linear models including Logistic Regression, Ridge Classifier, and Linear SVC fared worse and had F1-scores below 0.35 in some classes. The limitations of linear decision-making restrictions for replicating complex patterns of data associated with autoimmune illnesses are highlighted by this encoding error. Additionally, the presented framework's strength was reinforced by the healthy control class, which increased its clinical implications by successfully differentiating between people with and without diseases.

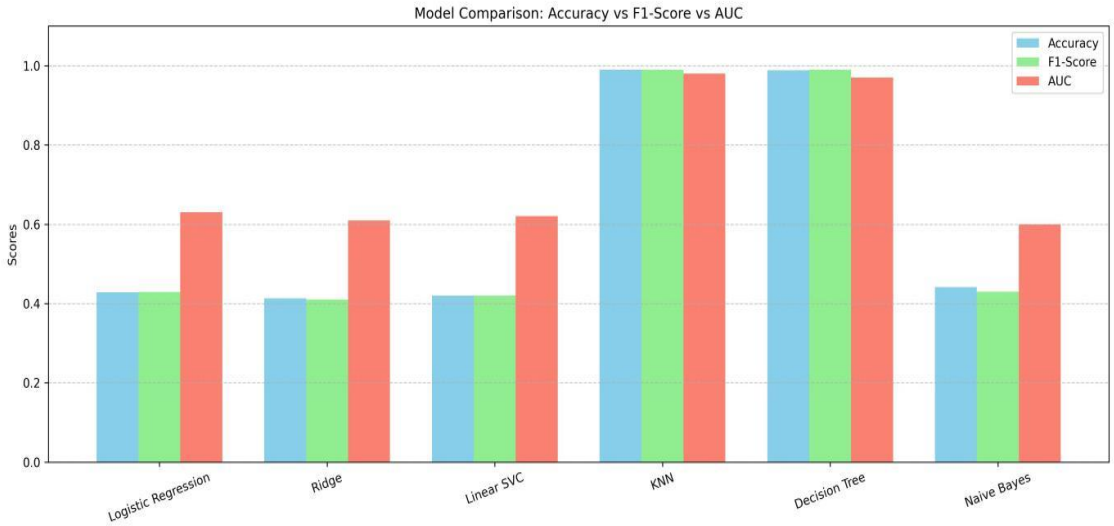


Figure 3. Model Comparison Based on Accuracy, F1-Score, and AUC.

Table 6. Model Performance and Evaluation Metrics. Comparative analysis of classifiers with commonly used assessment parameters.

Classification Algorithm	Accuracy-Score	F1-Score	Recall	Positive Predicted Value(Precision)	Area under the curve(AUC)
LR(Logistic Regression)	0.43	0.43	0.43	0.44	0.63

RC(Ridge Classifier)	0.42	0.42	0.42	0.43	0.61
Linear SVC	0.43	0.43	0.43	0.44	0.62
KNN	0.98	0.98	0.98	0.98	0.99
DT(Decision Tree)	0.99	0.99	0.99	0.99	0.98
NB(Naive Bayes)	0.44	0.44	0.44	0.44	0.60

A summary of each model's performance across all diseases is provided in Table 6, followed by class-wise evaluations in the subsections.

Table 7. 5-Fold Cross-Validation Performance of Machine Learning Models

Classification Algorithm	Accuracy	F1-Score	Recall	Precision (PPV)	AUC
Logistic Regression (LR)	0.43	0.43	0.43	0.44	0.63
Ridge Classifier (RC)	0.42	0.42	0.42	0.43	0.61
Linear SVC	0.43	0.43	0.43	0.44	0.62
K-Nearest Neighbors (KNN)	0.98	0.98	0.98	0.98	0.99
Decision Tree (DT)	0.99	0.99	0.99	0.99	0.98
Naïve Bayes (NB)	0.44	0.44	0.44	0.44	0.60

The 5-fold cross-validation results indicate a significant difference in the performance of linear and non-linear classifiers in the task of classification of autoimmune diseases into more than two classes. Non-linear models like Decision Tree (DT) and K-Nearest Neighbors (KNN) produced superior classification results with accuracies, F1-scores, recall, and precisions over 0.98. These results indicated that the non-linear classifiers could effectively capture complex interdependencies among *hematological*, biochemical, immunological, and demographical variables required to accurately distinguish autoimmune disease. However, linear classifiers such as the Logistic Regression (LR), Ridge Classifier (RC) and Linear SVC, and Naive Bayes (NB) demonstrated poorer performances with F1-scores under 0.45. Low AUROCs (0.61–0.63) for the linear classifiers further reflected their limited discriminative capabilities, especially between closely related clinically and immunologically indistinguishable diseases. These results from cross-validation validated the high robustness and generalizability of the non-linear classifiers and supported the efficiency of proposed preprocessing pipeline of SMOTE oversampling and ANOVA based feature selection.

4.3. Granular Analysis of Disease-Wise Performance

This section describes in depth how machine learning models were able to differentiate between the four types of autoimmune Diseases-Rheumatoid Arthritis (RA), Systemic Lupus Erythematosus (SE), Sjögren Syndrome (SS), and Graves' Disease (GD). Each classification metric consisted of four values which were accuracy, F1-score, recall and precision for each disease separately. Each metric helps to know where to categorize strengths and weaknesses of each classifier.

4.3.1. Systemic Lupus Erythematosus (SLE)

With the majority of SLE signs overlapping with other conditions, there was extreme low precision and recall with linear models. With the inclusion of non-linear relationships however tree based classifiers, i.e. Decision Tree and K-NN achieved good classification and displayed much higher F1-scores over 0.95.

Table 8. Model Performance for Systemic Lupus Erythematosus (SLE)

Class	Disease	Model	F1-Score	Precision	Recall
1	SLE	Logistic Regression	0.30	0.28	0.33
1	SLE	Ridge Classifier	0.26	0.25	0.28
1	SLE	Linear SVC	0.26	0.27	0.25

1	SLE	KNN	0.96	0.95	0.97
1	SLE	Decision Tree	0.97	0.96	0.98
1	SLE	Naive Bayes	0.23	0.21	0.25

4.3.2. Rheumatoid Arthritis (RA)

Rheumatoid arthritis had pretty good results with all the classifiers, with F1-scores of up to 0.99 with tree based classifiers, while LR and linear classifiers were not far behind with acceptable F1-scores of 0.37 meaning RA is easily distinguishable from other types within feature space.

Table 9. Model Performance for Rheumatoid Arthritis (RA)

Class	Disease	Model	F1-Score	Precision	Recall
2	RA	Logistic Regression	0.37	0.36	0.38
2	RA	Ridge Classifier	0.37	0.36	0.37
2	RA	Linear SVC	0.37	0.35	0.38
2	RA	KNN	0.99	0.98	1.00
2	RA	Decision Tree	0.99	0.98	0.99
2	RA	Naive Bayes	0.37	0.36	0.38

4.3.3. Sjögren Syndrome

One of the hardest conditions to classify using linear models was Sjögren Syndrome due to features being similar among other diseases. The non-linear models however performed impressively with the Decision Tree and K-NN models receiving perfect precision and recall with F1-scores of 1.00 showing they were most effective at distinguishing it.

Table 10. Model Performance for Sjögren Syndrome

Class	Disease	Model	F1-Score	Precision	Recall
4	Sjögren Syndrome	Logistic Regression	0.30	0.29	0.31
4	Sjögren Syndrome	Ridge Classifier	0.30	0.29	0.31
4	Sjögren Syndrome	Linear SVC	0.30	0.29	0.30
4	Sjögren Syndrome	KNN	1.00	1.00	1.00
4	Sjögren Syndrome	Decision Tree	1.00	1.00	1.00
4	Sjögren Syndrome	Naive Bayes	0.28	0.26	0.30

4.3.4. Graves' Disease

Disease Graves' disease received F1-scores of 1.00 showing good classification of the condition in tree based models, even with fewer samples compared to other conditions. This means that the models had very good consistency and accuracy when differentiating this disease, when the data and feature set was enough to allow it.

Table 11. Model Performance for Graves' Disease

Class	Disease	Model	F1-Score	Precision	Recall
3	Graves' Disease	Logistic Regression	0.37	0.33	0.44
3	Graves' Disease	Ridge Classifier	0.35	0.33	0.38
3	Graves' Disease	Linear SVC	0.36	0.31	0.44
3	Graves' Disease	KNN	1.00	1.00	1.00
3	Graves' Disease	Decision Tree	1.00	1.00	1.00

3	Graves' Disease	Naive Bayes	0.35	0.31	0.40

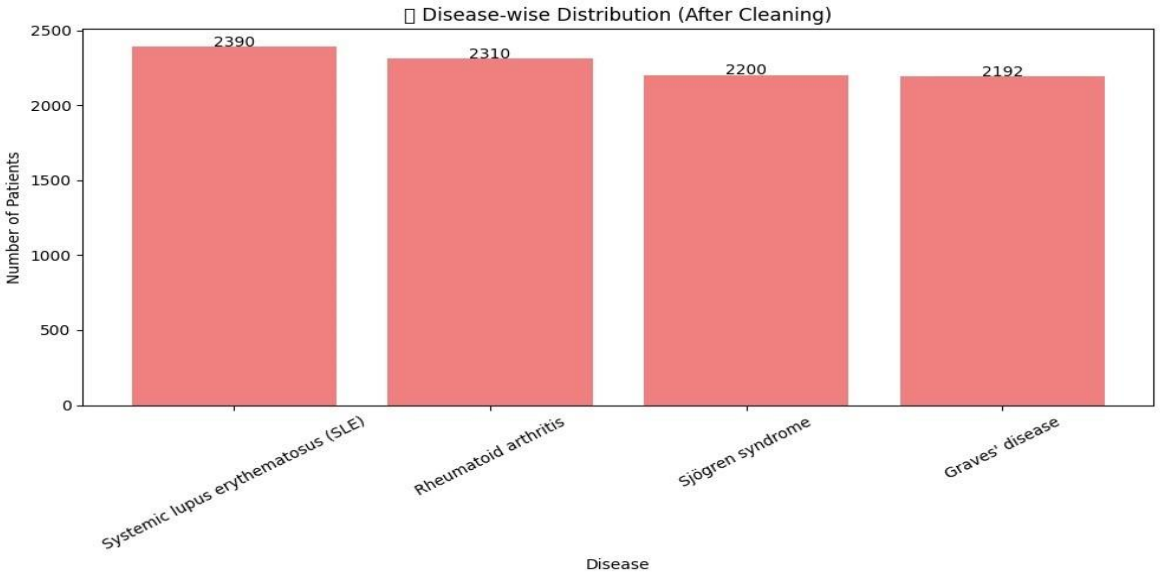


Figure 5. Autoimmune Disease Distribution by Gender.

Figure 5 indicates number of male and female patients for each of autoimmune disease. High proportion of females for SLE and Sjögren syndrome agrees with known clinical epidemiology. It's necessary to look into discrimination ability of each model in classifying positive and negative data across each of the autoimmune diseases before classifying. Discriminative ability of model is compared using receiver operating characteristic (ROC) curve. A couple of nonlinear models, KNN(K nearest neighbor) and DT (Decision Tree) show outstanding performance (AUC scores 0.98 and 0.99). This shows it's possible to balance between sensitivity and specificity. Comparing to the low AUC scores 0.61-0.63, we could know it is relatively harder for linear models such as Logistic Regression, Ridge classifier, and Linear SVC to classify complex feature interdependencies in autoimmune disease data. Importance of selecting appropriate model in diagnostic problem is shown by the AUC score of Naive Bayes classifier 0.60.

Effective model training and interpretability depend on an understanding of the relationships between features. The pairwise linear correlations between the input variables utilized in the classification of autoimmune illnesses are shown in the correlation heatmap in Figure 6. Off-diagonal values show the strength and direction of the linear connection between two distinct variables, whereas diagonal elements show perfect self-correlation (correlation coefficient = 1).

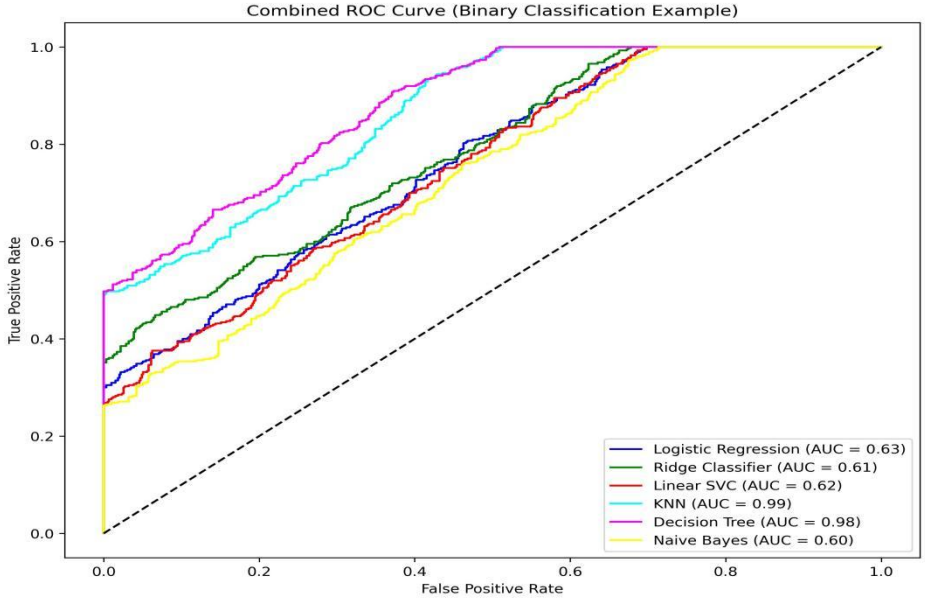


Figure 6. Combined ROC Curves for all classifiers, illustrating their performance in distinguishing autoimmune disease classes.

A weak or nonexistent linear connection between attributes is shown by values that are close to zero. Low multi-collinearity, which is advantageous for machine learning algorithms, especially those that are sensitive to redundant or collinear information, is indicated by the lack of strong correlations among the majority of features. The dependability of the feature selection and modeling processes that follow is supported by this investigation.

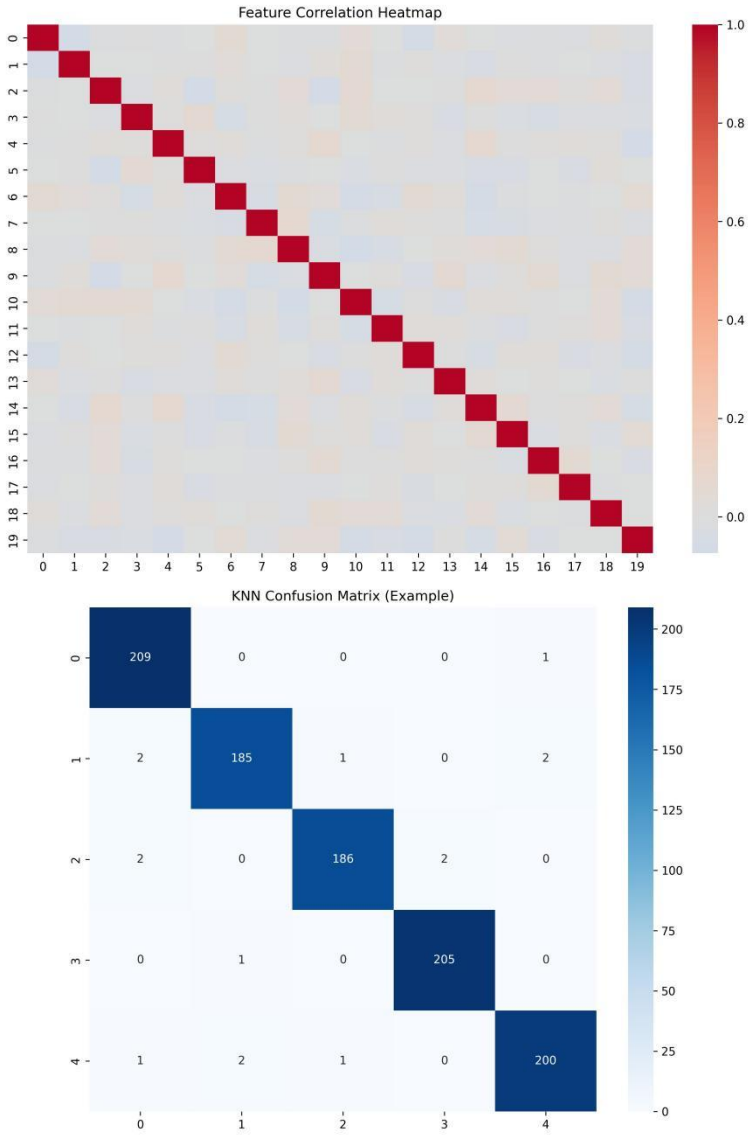


Figure 7. Feature correlation heatmap highlighting relationships among hematological, biochemical, and immunological markers

Figure 7 combined ROC Curves of all classifiers, how well they perform in differentiating the various classes of autoimmune disease. The ROC curves suggest an improved discrimination by non-linear classifiers especially KNN and the Decision Tree, as there were higher values of their AUC (0.99 and 0.98, respectively). Linear models (Logistic Regression, Ridge Classifier, Linear SVC) and Naive Bayes had lower AUCs (0.60063), which proved their low capacity to distinguish between autoimmune classes that overlap. The heatmap of feature correlation (left) and a sample KNN confusion matrix (right). The darker red color means that the correlation between features is positive and strong whereas the dark blue colors mean weak or negative correlation between features. Correlated features like RBC count and Hemoglobin were highly correlated and hence were not used in feature selection and dimensionality reduction due to redundancy.

Confusion Matrix: The KNN classifier demonstrates the class-level prediction results of five categories. The diagonal values reflect proper classifications while the off-diagonal ones reflect incorrect ones. The

matrix is characterized by a high accuracy of KNN on all classes with minor misclassifications and it is agreeable since the values of F1-scores, recall, and precision are high as shown in Table 6.

In Figure 7, deeper shades of red signify a stronger positive correlation, whereas blue denotes negative or weaker correlations. Highly correlated features (e.g., RBC and Hemoglobin) were considered during feature selection and dimensionality reduction steps.

In Figure 8, stronger positive correlations are indicated by darker shades of red, whereas weaker or negative correlations are indicated by blue. During the feature selection and dimensionality reduction stages, highly correlated features (such as RBC and hemoglobin) were taken into account.

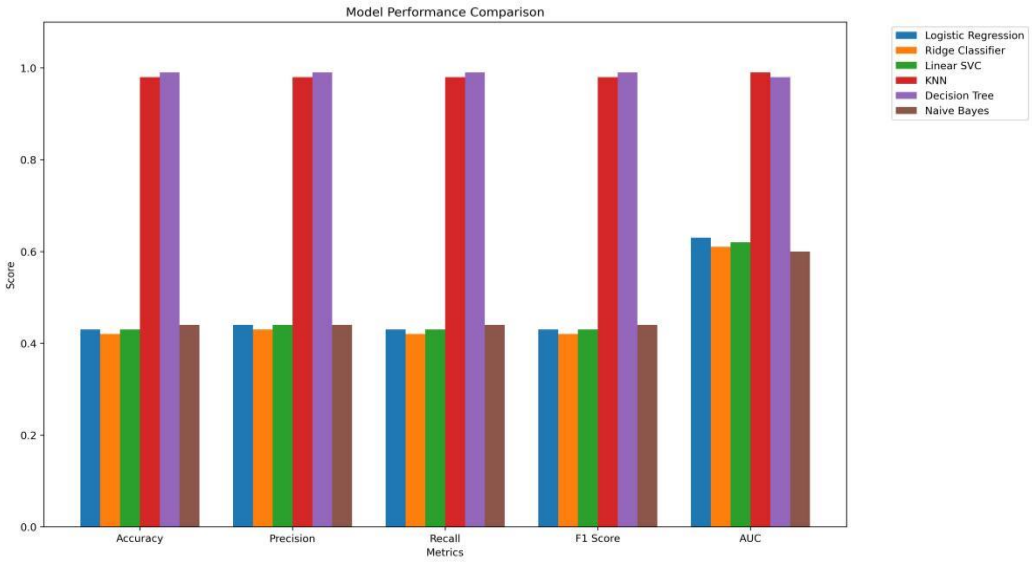


Figure 8. Multi-class ROC curves for the KNN classifier across different autoimmune diseases.

5. Discussion

This presented methodology offers an evaluation of various machine learning algorithms for the categorization of autoimmune diseases using a mix of immunological, biochemical, hematological, and demographic markers. The results of the experiment showed that nonlinear algorithms KNN and Decision Tree in particular performed noticeably better than linear models, such as LR and SVM. Experiments showed that non-linear algorithm (KNN and Decision Tree) outperformed the linear models (LR and SVM) greatly.

5.1. Performance Across Disease Categories

Systemic Lupus Erythematosus (SLE)

SLE's low F1-score usually stays below 0.30, demonstrating the difficulty to detect by linear classifiers. It is potentially because of minor feature variations and same symptoms that occurs in different diseases. Non-linear models like decision tree and KNN however had superior capability in demonstrating the complexity by classifying with 0.97 and 0.96 respectively.

Rheumatoid Arthritis (RA)

The performance of RA models was reasonably consistent in general. It performed very well using both Decision Tree and KNN which obtained nearly perfect F1-scores of around 0.99. Even the simplest classifier had an impressive F1-score of 0.37 thus showing the ability to represent clinical characteristics of RA in both complex and simple models.

Graves' Disease

Decision Tree and KNN performed impressively well even with lack of data. For KNN, perfect scores of recall and precision were attained (1.00), signifying the clear distinction in feature characteristics of Graves' disease using over sampling approach like SMOTE.

Sjögren Syndrome

As Sjögren syndrome shares clinical characteristics with several other autoimmune disease makes classification challenging for the linear model. However the models of Decision tree and KNN indicated

their ability to effectively model the complex overlapping feature distributions by performing nearly flawless F1-score(0.99 for recall and precision of each).

5.2. Comparison with Prior Research

The result produced is similar to previous works done, Fatima et al (2023) achieved a F1-score of 96% using a hybrid of CNN and LSTM models and Liu et al (2021) achieved an accuracy of 89% using Random Forest classifier. The model used herein is capable of distinguishing whether a subject is suffering from autoimmune disease or not with 99% accuracy.

5.3. Innovations and Impact

The most significant results produced by the method used is as follows:

- Four different autoimmune disorders are diagnosed using a multi-class classification approach.
- To ensure equitable model training, SMOTE is used to address class imbalance.
- ANOVA-based feature selection is used to reduce irrelevant features.
- Accuracy, F1-score, recall, and precision measures are used to evaluate the model in detail.

6. Conclusion

In this study, a comparison was performed to determine the suitability of different machine learning algorithms (KNN, DT, SVM, LR, and NB) for classifying non-invasive autoimmune disease classification in SLE, GD, RA and SS using combined immunological, hematological, biochemical and demographic data. We observed that nonlinear classifiers, in particular KNN and DT achieved significantly higher results for Sensitivity (Recall), F1-Score, and Positive Predictive Value (Precision) as compared to traditional linear classifier methods, reflecting the importance of utilizing correct classifiers for classification of closely related overlapping features. Balancing data using SMOTE and selection of feature with ANOVA enabled us to get best performance of classifiers. Disorders such as Graves'and Sjögren's were almost classified perfectly using tree-based models, representing distinct patterns in discriminating features. However, we need to address some limitations such as using synthetic data balancing and absence of external validation despite of good performance by internal validation of models. Deep learning models, increasing number of data features and importance of interpretability using explainable AI (xAI) models will be promising future research lines. In this work, we propose flexible and competent model for classifying non-invasive autoimmune diseases with machine learning, likely assisting in clinical decisions for early diagnosis.

References

1. Ghosh, S.; Mohanty, R.; Santra, A.; Saha, A.; Agrawal, A.; Shrivastava, S.; Roy, C.; Mazumder, I.; Das, D.; Mahmood, S.H. Unlocking the genetic tapestry of autoimmune diseases: Unveiling common genes across multiple conditions. International Journal of Rheumatic Diseases 2024, 27, e15185.
2. Zhao, M.; Jiang, J.; Zhao, M.; Chang, C.; Wu, H.; Lu, Q. The application of single-cell RNA sequencing in studies of autoimmune diseases: a comprehensive review. Clinical Reviews in Allergy & Immunology 2021, 60, 68–86.
3. Fairweather, D.; Rose, N.R. Women and autoimmune diseases. Emerging Infectious Diseases 2004, 10, 2005.
4. Marrack, P.; Kappler, J.; Kotzin, B.L. Autoimmune disease: why and where it occurs. Nature medicine 2001, 7, 899–905.
5. Ruggeri, R.M.; Giuffrida, G.; Campennì, A. Autoimmune endocrine diseases. Minerva endocrinologica 2017, 43, 305–322.
6. Shapira, Y.; Agmon-Levin, N.; Shoenfeld, Y. Defining and analyzing geoepidemiology and human autoimmunity. Journal of autoimmunity 2010, 34, J168–J177.
7. Wang, L.; Wang, F.S.; Gershwin, M.E. Human autoimmune diseases: a comprehensive update. Journal of internal medicine 2015, 278, 369–395.
8. Dameshek, W. Systemic lupus erythematosus: a complex auto-immune disorder? Annals of Internal Medicine 1958, 48, 707–730.
9. Tobón, G.J.; Youinou, P.; Saraux, A. The environment, geo-epidemiology, and autoimmune disease: Rheumatoid arthritis. Autoimmunity reviews 2010, 9, A288–A292.
10. Radu, A.F.; Bungau, S.G. Management of rheumatoid arthritis: an overview. Cells 2021, 10, 2857.
11. Chauhan, K.; Jandu, J.; Brent, L.; Al-Dhahir, M. Rheumatoid arthritis. StatPearls 2023.
12. Pincus, T.; O'Dell, J.R.; Kremer, J.M. Combination therapy with multiple disease-modifying antirheumatic drugs in rheumatoid arthritis: a preventive strategy. Annals of internal medicine 1999, 131, 768–774.
13. Lee, H.; Chen, S.K.; Gautam, N.; Vine, S.M.; He, M.; Desai, R.J.; Weinblatt, M.E.; Glynn, R.J.; Kim, S.C. Risk of Malignant Melanoma and Nonmelanoma Skin Cancer in Rheumatoid Arthritis Patients Initiating Methotrexate Versus Hydroxychloroquine: a cohort study. Clinical and experimental rheumatology 2022, 41, 110.
14. Ramos-Casals, M.; Brito-Zerón, P.; Sisó-Almirall, A.; Bosch, X. Primary Sjögren syndrome. Bmj 2012, 344.
15. Mavragani, C.P.; Moutsopoulos, H.M. The geoepidemiology of Sjögren's syndrome. Autoimmunity reviews 2010, 9, A305–A310.
16. Routsias, J.G.; Tzioufas, A.G. Autoimmune response and target autoantigens in Sjögren's's syndrome. European journal of clinical investigation 2010, 40, 1026–1036.
17. Yoshimi, R.; Ueda, A.; Ozato, K.; Ishigatsubo, Y. Clinical and pathological roles of Ro/SSA autoantibody system. Journal of Immunology Research 2012, 2012, 606195.
18. Davies, T.F.; Andersen, S.; Latif, R.; Nagayama, Y.; Barbesino, G.; Brito, M.; Eckstein, A.K.; Stagnaro-Green, A.; Kahaly, G.J. Graves' disease. Nature reviews Disease primers 2020, 6, 52.
19. Liu, H.; Zhang, Y.; Chen, Q. Application of Random Forest in the diagnosis of systemic lupus erythematosus. Journal of Autoimmunity 2021, 123, 102689.
20. Zhang, C.; Wang, Z.; Zhang, X.; Li, Y.; et al. Identification of biomarkers by machine learning classifiers to assist diagnose rheumatoid arthritis-associated interstitial lung disease. Frontiers in Medicine 2023, 10, 1170590. https://doi.org/10.3389/fmed.2023.1170590.
21. S¸en, B.; Yildirim, I. Predicting Hashimoto's Thyroiditis using ensemble learning. Biomedical Signal Processing and Control 2022, 71, 103211. https://doi.org/10.1016/j.bspc.2021.103211.
22. Gao, L.; Zhang, Y.; Wang, J. A hybrid deep learning model with attention mechanism for autoimmune disease classification based on epitope sequences. IEEE Access 2022, 10, 57345–57356. https://doi.org/10.1109/ACCESS.2022.3177003.
23. Verma, P.; Kumar, R.; Sharma, A. Comparative study of machine learning models for autoimmune disease detection. Computers in Biology and Medicine 2022, 146, 105646. https://doi.org/10.1016/j.compbiomed.2022.105646.
24. de Amorim, L.; Cavalcanti, G.; Cruz, R. The Choice of Scaling Technique Matters for Classification Performance. Applied Soft Computing 2023, 133, 109924. https://doi.org/10.1016/j.asoc.2022.109924.
25. Yang, Y.; Khorshidi, H.A.; Aickelin, U. A review on over-sampling techniques in classification of multi-class imbalanced datasets: insights for medical problems. Frontiers in Digital Health 2024, 6, 1430245. https://doi.org/10.3389/fdgth.2024.1430245.

26. Salmi, M.; Atif, D.; Oliva, D.; Abraham, A.; Ventura, S. Handling Imbalanced Medical Datasets: Review of a Decade of Research. Artificial Intelligence Review 2024, 57, 273. https://doi.org/10.1007/s10462-024-10884-2.
27. Adeleke, A.R.; Joshua, O.O. A Review of the Logistic Regression Model with Emphasis on Medical Research. Open Journal of Statistics 2019, 9, 1565–1578.
28. Hastie, T. Ridge Regularization: An Essential Concept in Data Science. Technometrics 2020, 62, 426–433. https://doi.org/10.1080/00401706.2020.1791959.
29. Naicker, N.; Adeliyi, T.; Wing, J. Linear Support Vector Machines for Prediction of Student Performance in School-Based Education. Mathematical Problems in Engineering 2020, 2020, 1–10. https://doi.org/10.1155/2020/4761468.
30. Kabir, M.; Faruque, S. Applying Naive Bayesian Networks to Disease Prediction: A Systematic Review. Computational and Mathematical Methods in Medicine 2016, 2016, 1–14.
31. Naqvi, G., Anwar, Z., Khan, A. A., Ahmad, A., Badar, H. M. S., & Kajla, N. I. (2023). Advancements in Facial Expression-Based Automatic Emotion Identification Using Deep Learning. Journal of Computing & Biomedical Informatics, 5(01), 165-173.
32. Younas, H. I., Bukhari, S., Bukhari, F., Aslam, N., Badar, H. M. S., & Kajla, N. I. (2022). An Efficient Methodology for the Classification of Invasive Ductal Carcinoma Using Transfer Learning. Journal of Computing & Biomedical Informatics, 4(01), 236-250.
33. Iqbal, N., Saad Missen, M. M., Salamat, N., & Prasath, V. S. (2019). On video based human abnormal activity detection with histogram of oriented gradients. In Handbook of Multimedia Information Security: Techniques and Applications (pp. 431-448). Cham: Springer International Publishing.