Jesse Byers

Random Forest Classification of Rheumatic and Autoimmune Diseases

Agenda

- Introduction
- Context of Study
- Data Analysis Process
- Key Insights
- Limitations of Study
- Recommended Course of Action
- Expected Benefits of the Study

Introduction

- Jesse Byers
  - Master of Science in Data Analysis – Data Science student
  - 20 years in Education and Non-Profits
  - 5 years in Web Development and Data Analytics
  - Parent of a child with a rheumatic autoimmune disease

Executive Summary

Goal:

- Develop a predictive model to support diagnosis of Rheumatic and Autoimmune Diseases based on blood test results

Data Analysis Process

- Addressing Missing Data Values
- Random Forest Classification Model
- Permutation Testing

Key Findings and Limitations

- Predictive modeling shows promise, but not yet ready to be used in clinical practice
- Model performance limited by unbalanced data set and missing values

Recommended Course of Action

- Refine predictor variables, explore other imputation techniques, and generate a more balanced dataset

Context of Study

Rheumatic and autoimmune diseases can be very challenging to diagnose and treat:

Overlapping, nonspecific symptoms → Delays in getting a diagnosis → Delays in starting treatment

If a predictive diagnostic model is successful:

Faster diagnostic time → Earlier treatment → Better patient outcomes → Reduced health care costs

Research Question and Hypotheses

To what extent do blood test result variables predict a specific autoimmune disease diagnosis?

- Null Hypothesis:
  - Blood test result variables alone can not accurately predict a disease
- Alternate Hypothesis:
  - Blood test result variables statistically significantly predict a disease.

Data Analysis Process

- Collection of Data
- Addressing Missing Data Values
- Exploratory Data Analysis
- Random Forest Classification Model
- Permutation Testing

Collection of Data

Diagnosis of Rheumatic and Autoimmune Diseases Dataset

https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/VM4OR3

Disease Label	# of Samples
Rheumatoid Arthritis	2,848
Reactive Arthritis	516
Ankylosing Spondylitis	2,127
Sjogren's Syndrome	1,852
Systemic Lupus Erythematosus	1,355
Psoriatic Arthritis	1,783
Normal (no disease)	1,604

Feature	Data Type
Age	Numeric (continuous)
Sex	Categorical
ESR	Numeric (continuous)
CRP	Numeric (continuous)
RF	Numeric (continuous)
Anti-CCP	Numeric (continuous)
HLA-B27	Categorical
ANA	Categorical
Anti-Ro	Categorical
Anti-La	Categorical
Anti-dsDNA	Categorical
Anti-Sm	Categorical
C3	Numeric (continuous)
C4	Numeric (continuous)

Addressing Missing Data Values

- Data is Missing Not At Random (MNAR)

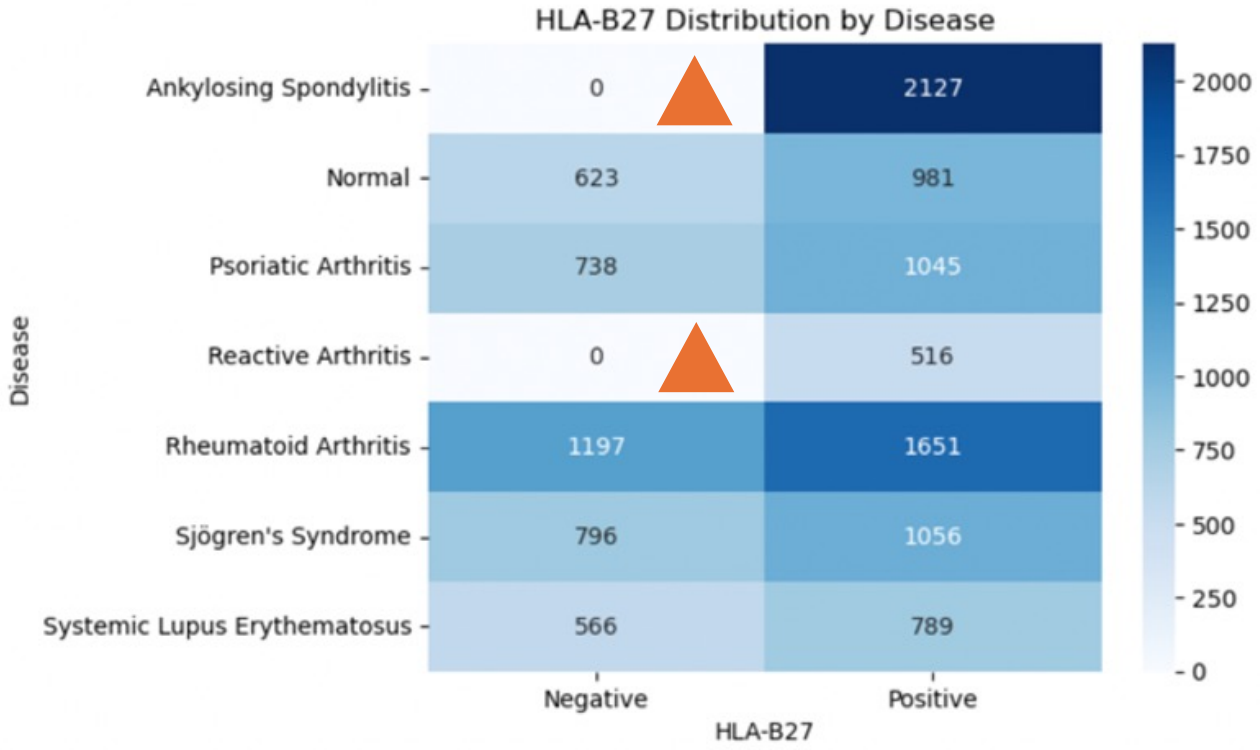
- 3 Strategies Explored and Rejected
  - Delete all rows with missing values
  - Impute with normal value
  - Impute with mean of similar patients

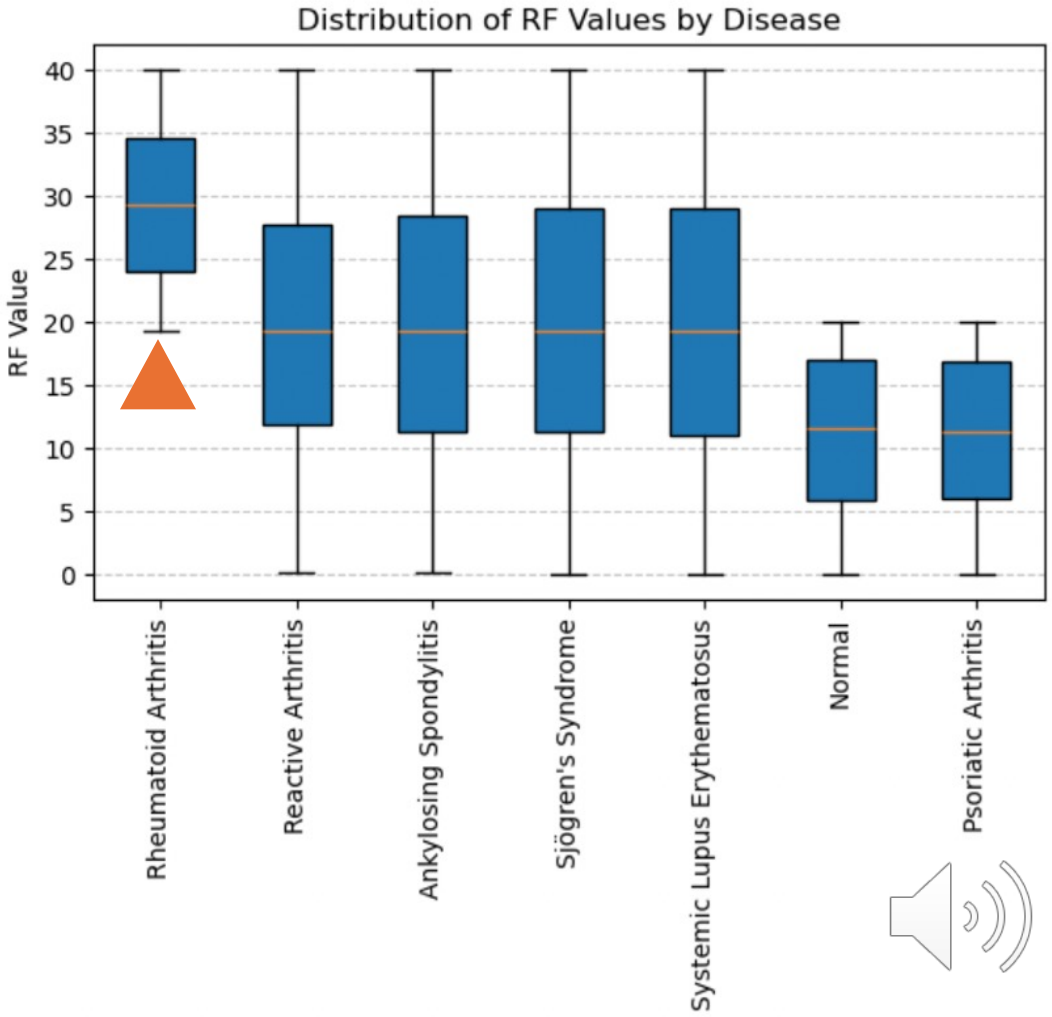
- Implemented Strategy
  - Use SimpleImputer to impute mean values
  - Add an indicator flag for missingness

Feature	# Missing	% Missing
ESR	1088	9%
CRP	2417	20%
RF	1329	11%
Anti-CCP	3263	27%
HLA-B27	1934	16%
ANA	3746	31%
Anti-Ro	2900	24%
Anti-La	3021	25%
Anti-dsDNA	4713	39%
Anti-Sm	5197	43%
C3	1692	14%
C4	2054	17%

Exploratory Data Analysis

The distribution of all predictor variables were analyzed by disease:





Exploratory Data Analysis

Feature(s)	Strongest Disease Associations
High ESR and CRP	Rheumatoid Arthritis Reactive Arthritis Ankylosing Spondylitis Psoriatic Arthritis
High RF and Anti-CCP	Rheumatoid Arthritis
Low C3 and C4	Systemic Lupus Erythematosus
Positive HLA-B27	Ankylosing Spondylitis Reactive Arthritis
Negative ANA	Sjogren's Syndrome Systemic Lupus Erythematosus
Positive Anti-Ro and Anti-La	Sjogren's Syndrome
Positive Anti-dsDNA and Anti-Sm	Systemic Lupus Erythematosus

Random Forest Classification Model

Ensemble tree-based learning approach

Predictor features are run through multiple decision trees

Majority voting is used to assign the final prediction

*"The RF classifier is **one of the most successfully implemented ensemble learning techniques** which have proved very popular and powerful for high-dimensional classification and skewed problems in pattern recognition and ML.*

*It offers the benefit of computing efficiency and **improves the accuracy of predictions without considerably increasing calculation costs**.*

*Based on these characteristics, most of the researchers recorded the **highest value of accuracy for the prediction of diseases**."*

- Ray & Chaudhuri, 2021

Random Forest Classification Model

Final Data Structure

- Missingness indicators
- One-Hot Encoding: Predictor Features
- Label Encoding: Target Feature

Split Dataset

- Unbalanced Representation
  - Reactive Arthritis (low)
  - Rheumatoid Arthritis (high)

Out[5]:

	Age	Disease	ESR	CRP	RF	Anti-CCP	C3	C4	ESR_missing	CRP_missing	...	ANA_Negative	ANA_Positive
0	70	4	39.0	18.6	34.2	29.9	133.0	27.0	0.0	0.0	...	1	0
1	39	4	26.0	21.7	35.5	28.9	100.0	66.0	0.0	0.0	...	0	1
2	36	4	41.0	15.6	21.3	21.3	158.0	12.0	0.0	0.0	...	1	0
3	35	4	43.0	23.4	26.0	39.0	119.0	41.0	0.0	0.0	...	0	1
4	37	4	30.0	15.6	38.1	30.8	144.0	49.0	0.0	1.0	...	1	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
12080	32	2	36.0	17.0	14.5	16.1	133.0	32.0	0.0	0.0	...	0	1
12081	36	2	43.0	15.6	17.7	13.5	133.0	41.0	0.0	1.0	...	1	0
12082	20	2	31.0	28.8	4.8	5.8	133.0	38.0	0.0	0.0	...	0	1
12083	33	2	36.0	15.6	19.2	9.5	96.0	52.0	0.0	1.0	...	0	1
12084	48	2	32.0	26.9	7.2	13.6	178.0	38.0	0.0	0.0	...	0	1

12085 rows × 34 columns

60% train	20% validate	20% test
• 303 – 1747 samples per disease class	• 112 – 522 samples per disease class	• 101 – 579 samples per disease class

Random Forest Classification Model

Initial Model	• Scikit-Learn RandomForestClassifier using default parameter values • **Accuracy of 83.33%**
Hyperparameter Tuning	• Scikit-Learn RandomizedSearchCV
Tuned Model	• Scikit-Learn RandomForestClassifier using best estimators • **Improved accuracy of 83.78%**

Permutation Testing

Justification

Evaluate how the model performs compared to making predictions by random chance

Implemented through Scikit-Learn permutation_test_score

Creates permutations of samples with randomly shuffled disease labels

Results

Accuracy: 83%

P-value: 0.0099

There is less than 1% chance that the model's predictive power could be explained by chance.

Key Finding 1:

Analysis supports the alternate hypothesis.

This is very strong evidence that the set of predictor variables, which include only objective blood test results, can statistically significantly predict the target labels for rheumatic and autoimmune diseases.

- The model can predict an accurate diagnosis across 7 disease classes about 84% of the time.
- There is less than 1% chance that the model's predictive power could be explained by chance.

Key Finding 2:

The model is not yet ready to support clinical diagnosis.

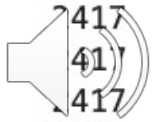
The model makes highly accurate predictions on some diseases, but less accurate performance on others.

- Overall – 84%
- Ankylosing Spondylitis – 65%
- Systemic Lupus Erythematosus – 98%

```
Tuned Model Accuracy: 83.78%
Tuned Model ROC-AUC score: 98.42%
Tuned Model Classification Report
                          precision    recall  f1-score   support

    Ankylosing Spondylitis       0.69      0.61      0.65       413
                    Normal       0.94      0.74      0.83       301
     Psoriatic Arthritis         0.86      0.89      0.88       343
     Reactive Arthritis          0.92      0.54      0.68       101
   Rheumatoid Arthritis          0.80      0.93      0.86       579
     Sjögren's Syndrome          0.83      0.96      0.89       404
Systemic Lupus Erythematosus     1.00      0.97      0.98       276

               accuracy                           0.84      2417
              macro avg          0.86      0.80      0.82      2417
           weighted avg          0.84      0.84      0.83      2417
```

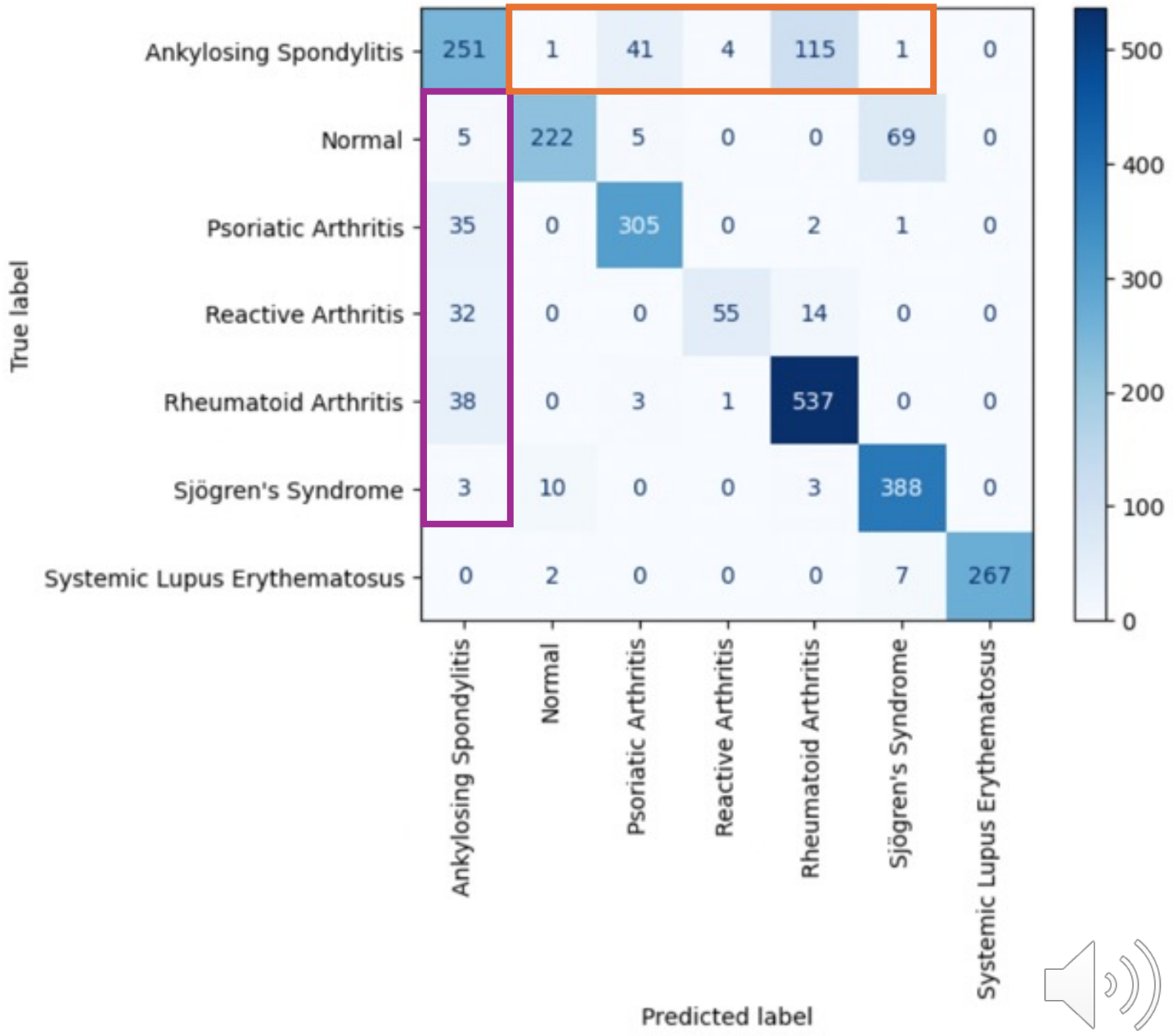


The model has trouble differentiating between different types of Arthritis.

For Ankylosing Spondylitis patients:

- **False negatives** with predicted diagnoses of Rheumatoid Arthritis and Psoriatic Arthritis.

- **False positives**, misdiagnosing patients who have Psoriatic, Rheumatoid, and Reactive Arthritis with Ankylosing Spondylitis instead.



Limitations of the Predictive Model

Model performance was limited by unbalanced disease classes, and volume of missing values.

*"Generally, **healthcare datasets are highly imbalanced** in nature. Classification of these datasets results in erroneous prediction and inaccurate accuracies. Another very common characteristic of healthcare datasets is **a large number of missing data values** for multiple features."*

- Ray & Chaudhuri, 2021

Recommended Course of Action:

Engage in further experimentation with the model towards a goal of 85% accuracy across each individual disease class

Refine and/or redefine the predictor variables

- Convert numerical to categorical
- Add columns for normal/abnormal flags

Use a more complex imputation technique, or a tool that natively addresses missing values

- Multiple Imputation technique
- Gradient Boosting model or XG Boost model

Generate a more robust dataset for analysis

- Represent broader population from multiple settings (countries, clinical settings)
- Include greater diversity of demographic features (age, gender and race)
- Increase number of samples from rare disease cla[illegible]s to balance the classe[illegible]

Expected Benefits of Study

If a predictive diagnostic model is successful:

- Faster diagnostic time
- Earlier treatment
- Better patient outcomes
- Reduced health care costs

Autoimmune diseases are becoming more prevalent and can be very costly to treat (Miller, 2023).

Diagnosis using objective features, such as blood tests, has the most potential to benefit women and patients from underserved communities (Myers, 2025).

Recap: Executive Summary

Goal:

- Develop a predictive model to support diagnosis of Rheumatic and Autoimmune Diseases based on blood test results

Data Analysis Process

- Addressing Missing Data Values
- Random Forest Classification Model
- Permutation Testing

Key Findings and Limitations

- Predictive modeling shows promise, but not yet ready to be used in clinical practice
- Model performance limited by unbalanced data set and missing values

Recommended Course of Action

- Refine predictor variables, explore other imputation techniques, and generate a more balanced dataset

To Learn More

See the full report, **Random Forest Classification of Rheumatic and Autoimmune Diseases,** by Jesse Byers

- The following appendices are included with the report:
  - Appendix 1: Excel file with original dataset
  - Appendix 2: Jupyter notebook with missing data experimentation
  - Appendix 3: Cleaned and imputed dataset
  - Appendix 4: Jupyter notebook with model code and evaluation
  - Appendix 5-7: Training, validation, and testing datasets

References

Ashtari, H. (2024). *XGBoost vs. Random Forest vs. Gradient Boosting: Differences | Spiceworks*. Spiceworks. https://www.spiceworks.com/tech/artificial-intelligence/articles/xgboost-vs-random-forest-vs-gradient-boosting/

Austin, P. C., White, I. R., Lee, D. S., & van Buuren, S. (2020). Missing Data in Clinical research: a Tutorial on Multiple Imputation. *Canadian Journal of Cardiology*, *37*(9). https://doi.org/10.1016/j.cjca.2020.11.010

Creative Commons. (2019). *Creative Commons — CC0 1.0 Universal*. Creativecommons.org.

https://creativecommons.org/publicdomain/zero/1.0/

Dhafar Hamed Abd; Mohammed Fadhil Mahdi; Arezoo Jahani. (2025). "Rheumatic and autoimmune diseases dataset", https://doi.org/10.7910/DVN/VM4OR3, Harvard Dataverse, V3

Mahdi, M. F., Jahani, A., & Abd, D. H. (2025). Diagnosis of rheumatic and autoimmune diseases dataset. *Data in Brief*, *60*, 111623. https://doi.org/10.1016/j.dib.2025.111623

Miller, F. W. (2023). The increasing prevalence of autoimmunity and autoimmune diseases: an urgent call to action for improved understanding, diagnosis, treatment, and prevention. *Current Opinion in Immunology*, *80*(102266), 102266. https://doi.org/10.1016/j.coi.2022.102266

Myers, Emma. (2025). *Silent Struggles: How Autoimmune Diseases in Women Are Overlooked –*

North Carolina Schweitzer Fellowship. Ncschweitzerfellowship.org. https://ncschweitzerfellowship.org/silent-struggles-how-autoimmune-diseases-in-women-are-overlooked/

Ray, A., & Chaudhuri, A. K. (2021). Smart healthcare disease diagnosis and patient management: Innovation, improvement and skill development. *Machine Learning with Applications*, *3*, 100011. https://doi.org/10.1016/j.mlwa.2020.100011

Scikit-Learn. (2025). *sklearn.ensemble.RandomForestClassifier — scikit-learn 0.20.3 Documentation*. Scikit-Learn.org. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html