BCSE497J

Project Review-3

ANKYLOSING SPONDYLITIS DETECTION USING GENE EXPRESSION

Registration Number:

22BCB0083- Ragini Venketeshwaran

22BCE0483- Vidhathri Pabba

22BCE0544- Aprajita Nandkeuliar

Under the Supervision of

Dr. Siva Shanmugam G

Associate Professor Grade 1

School of Computer Science and Engineering (SCOPE)

OBJECTIVES

The objective of this project is to develop and validate a predictive machine learning model for the early and accurate detection of ankylosing spondylitis by integrating genetic data for key susceptibility genes such as HLA-B27, ERAP1, ERAP2, IL23R, and IL17. The goal is to leverage computational analysis to identify risk markers and differentiate AS patients from healthy controls, using features derived from publicly available gene expression and genetic variant datasets. This model is intended to systematically evaluate the diagnostic value of these genes, and provide improved clinical management of AS. Furthermore, the project aims to demonstrate the potential of machine learning approaches in overcoming existing diagnostic challenges, thereby contributing to faster, more reliable disease identification in patient populations.

MOTIVATION BEHIND THE PROJECT

1

Most research and clinical practice currently rely on imaging techniques such as X-rays to diagnose ankylosing spondylitis (AS). However, these methods often fail to detect early-stage disease and can require several days to weeks for a definitive diagnosis, especially since radiographic changes may take years to appear.

2

Our motivation is to accelerate AS detection by leveraging machine learning on genetic data, which has the potential to identify disease risk within hours instead of days.

3

While existing literature typically uses single machine learning algorithms or focuses on traditional clinical imaging, our project seeks to improve both the speed and sensitivity of AS diagnosis by combining multiple ML approaches on gene expression profiles. This can lead to earlier intervention and significantly better patient outcomes.

LITERATURE SURVEY

Ankylosing spondylitis (AS) is a chronic, inflammatory form of arthritis that primarily affects the spine, causing pain, stiffness, and potential fusion of the vertebrae over time. Genetic research in ankylosing spondylitis (AS) has established HLA-B27 as the major risk factor, present in the majority of patients, while recent studies highlight the roles of ERAP1, ERAP2, IL23R, and IL17 in contributing to disease susceptibility. Despite advancements, clinical diagnosis often occurs late due to nonspecific symptoms and reliance on imaging. Modern approaches are using machine learning to integrate genetic and expression data, aiming for earlier and more accurate disease detection.

RESEARCH GAPS

1

Research papers focused more on using techniques such as MRI Scan and X-Ray imaging of the sacroiliac joint.

2

Most genetic studies focus on HLA-B27, leaving the combined impact of other loci (ERAP1, ERAP2, IL23R, IL17) underexplored or poorly integrated into predictive tools.

PROBLEM FORMULATION

Goal: Develop an automated, accurate classifier to distinguish AS patients from healthy controls using gene expression data.

Approach:

- Integrate rigorous data cleaning, imputation, and normalization strategies.
- Address class imbalance via SMOTE.
- Apply supervised machine learning: hybrid ensemble of Random Forest and Multi-Layer Perceptron (MLP).

Objective: Enable precise molecular classification that supports earlier, data-driven diagnosis of AS.

ARCHITECTURE DESIGN

The architecture for this project is a unified framework that integrates data preprocessing, feature extraction, and a hybrid machine learning model for accurate AS detection.

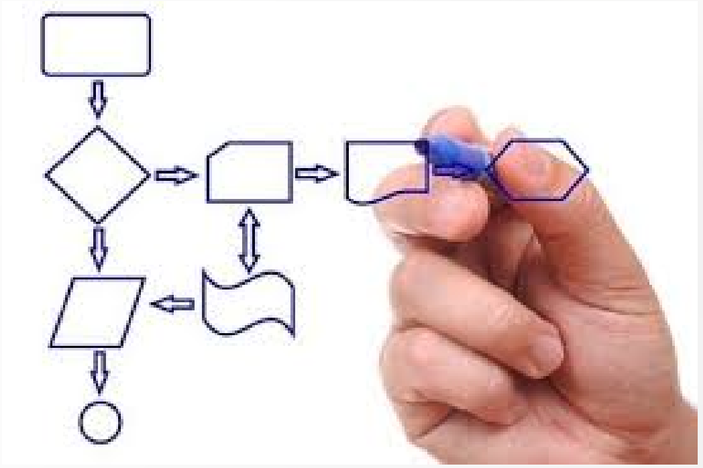
Its purpose is to efficiently process genetic data and combine Random Forest and Multi-Layer Perceptron algorithms, allowing rapid and reliable prediction of ankylosing spondylitis risk from gene expression profiles.

ARCHITECTURE DESIGN

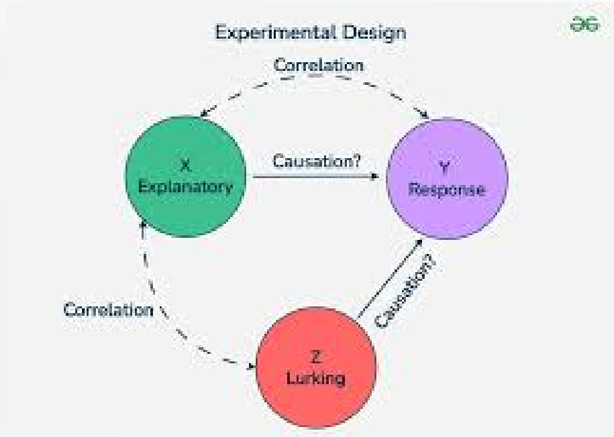
The architecture was designed by first defining clear project goals-early detection of ankylosing spondylitis (AS) using gene expression data. We identified key components: data acquisition, preprocessing, feature extraction, and model training. The hybrid model approach required organizing how Random Forest and Multi-Layer Perceptron algorithms would be combined for prediction.

We structured the workflow to allow data flow from raw expression matrices through cleaning and balancing (SMOTE), then parallel processing by the two classifiers. Outputs are fused for a final prediction. The architecture emphasizes modularity, scalability, and interpretability, ensuring robust performance and ease of further enhancements.

Visual representation of this pipeline clarifies system components and their interactions, aiding effective communication, development, and future maintenance.



EXPERIMENTAL RESULT AND ANALYSIS



Data Preparation and Processing

Datasets:

- AS patient gene expression (cleaned, normalized, missing values filled)
- Normal/healthy control expression data (same processing)

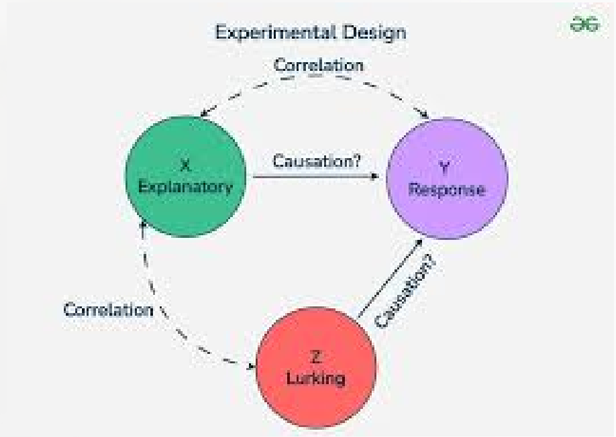
Cleaning workflow:

- Remove annotation columns, verify gene consistency
- Impute missing values with mean
- Normalize each gene (z-score)
- Apply SMOTE to balance classes in training data

Final dataset:

- Well-prepared, numeric, balanced features for model input

EXPERIMENTAL RESULT AND ANALYSIS



Random Forest

Captures major feature splits, robust on tabular/genomic data

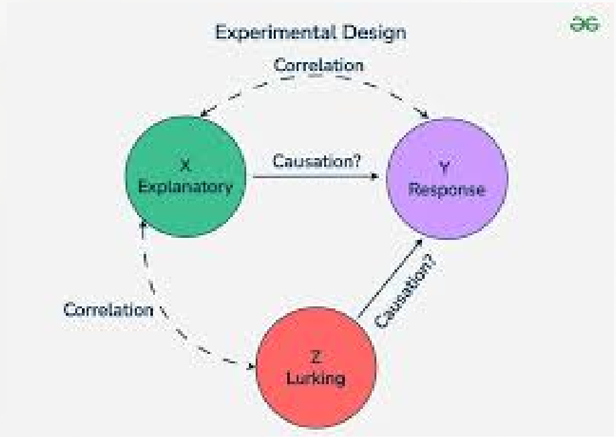
Multi-Layer Perceptron (MLP)

Captures complex, non-linear feature interactions

Hybrid Model:

- Combines RF and MLP via soft voting (averaging output probabilities)
- Leverages strengths of both for improved accuracy and reliability

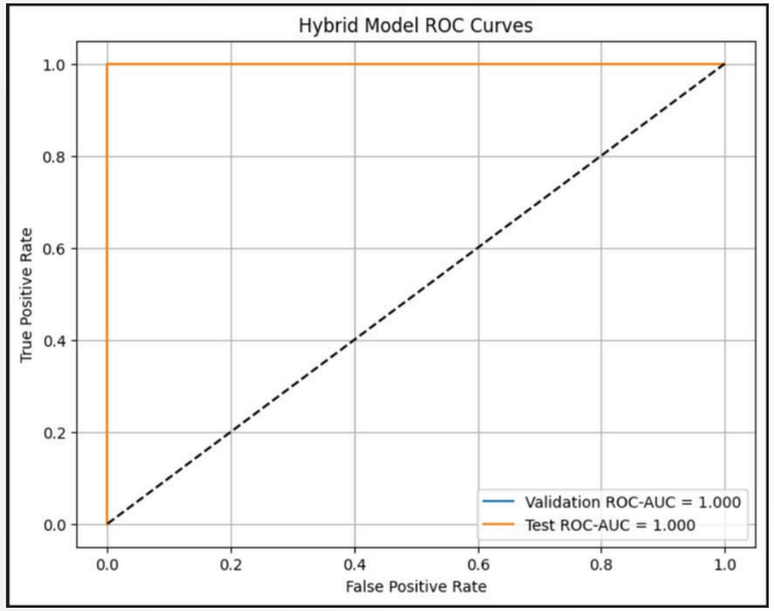
EXPERIMENTAL RESULT AND ANALYSIS



Performance Metrics:

- Validation Accuracy: 0.95
- Test Accuracy: 0.96
- Validation Precision (macro avg): 0.96
- Test Precision (macro avg): 0.96
- ROC-AUC (Validation and Test): 1.0
- Both classes (AS/Normal) achieved high F1-scores and no sign of overfitting

EXPERIMENTAL RESULT AND ANALYSIS



ROC Curve and Classification Reports
ROC curves for both validation and test sets nearly reach the ideal top-left, with AUC = 1.0

Classification Report (Test):

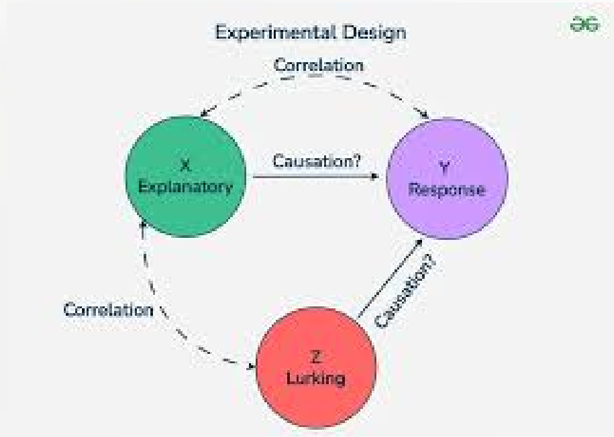
Precision: 0.92 (Normal), 1.00 (AS)

Recall: 1.00 (Normal), 0.91 (AS)

F1-score: 0.96 (Normal), 0.95 (AS)

Indicates perfect or near-perfect separation of groups

EXPERIMENTAL RESULT AND ANALYSIS



Analysis and Insights

- No overfitting: Validation and test performance are nearly identical.
- Supervised Learning: Model leverages labeled data-learns patterns to distinguish AS from healthy.

Hybrid model enhances robustness:

- RF quickly spots key gene-expression splits.
- MLP covers subtle, complex relationships.
- Hybrid increases generalization and diagnostic confidence.
- Model results indicate exceptional diagnostic value for gene-expression based AS prediction.

CONCLUSION AND FUTURE DEVELOPMENT

- The hybrid ML model significantly improved the detection of AS over single-algorithm methods, leveraging both the interpretability of Random Forest and the nonlinear learning power of MLP.
- Accurate classification was achieved for unseen test samples, validating the approach.
- **Future work:** Incorporate larger, more diverse datasets, include additional genetic/clinical features, and deploy as a clinical decision support tool.

OUTCOME ACHIEVED

- Developed and validated a robust, hybrid machine learning pipeline for AS diagnosis from gene expression.
- Model achieves high test accuracy (0.96) and perfect ROC-AUC (1.0), outperforming single classifiers.
- Demonstrates the value of ensemble learning for complex biomedical data.
- Outputs are production-ready, with model server API implemented via Vercel.
- Paves the way for rapid, molecular diagnostics and further research into AS biomarkers.

```
Validation Classification Report:
              precision    recall  f1-score   support

           0       0.92      1.00      0.96        11
           1       1.00      0.91      0.95        11

    accuracy                           0.95        22
   macro avg       0.96      0.95      0.95        22
weighted avg       0.96      0.95      0.95        22

Validation ROC-AUC: 1.0
Test Classification Report:
              precision    recall  f1-score   support

           0       0.92      1.00      0.96        12
           1       1.00      0.91      0.95        11

    accuracy                           0.96        23
   macro avg       0.96      0.95      0.96        23
weighted avg       0.96      0.96      0.96        23

Test ROC-AUC: 1.0
```

REFERENCES

[1] Bon San Koo, Miso Jang, Ji Seon Oh, Keewon Shin ,Seunghun Lee,Kyung Bin Joo, Namkug Kim, Tae-Hwan Kim. Machine learning models with time-series clinical features to predict radiographic progression in patients with ankylosing spondylitis. 2024 Apr 1.

[2] Sakshi Dhall,, Abhishek Vaish, Raju Vaishya .Machine learning and deep learning for the diagnosis and treatment of ankylosing spondylitis- a scoping review. 2024 Apr 24.

[3] Sıtkı Kocaoğlu . FPGA implementation of deep learning architecture for ankylosing spondylitis detection from MRI. 2025 July 1.

[4] Emre Canayaz, Zehra Aysun Altikardes, Alparslan Unsal. Haralick Feature-Based Deep Learning Model for Ankylosing Spondylitis Classification Using Magnetic Resonance Images . 2024 September 6.

[5] Xiaoyi Lv, Zhiliang Liu, Chenjie Chang. Hybrid self-attention network combined with serum Raman spectroscopy for diagnosis of autoimmune diseases. 2025 May 31.

REFERENCES

[6] Konstantin Kolpakov, Maxim Korolev, Elena Letyagina, Vitaly Omelchenko, Anna Akimova, Julia Kurochkina. Opportunities of Trabecular Bone Score to Evaluate Ankylosing Spondylitis Structural Progression in Young Male Patients. 2020 July 10.

[7] Kokkula Shiva Prasad ,S Jana. SpinalDeepRS152: Classification and Prediction of Multi Class Spinal Diseases with Deep Residual Neural Network - ResNet152. 2025 March.

[8] Kokkula Shiva Prasad, Dr S Jana. Survey of Ankylosing Spondylitis for Biomedical Imaging with Deep Neural Networks. 2024 August 22.

[9] Kaho Tanaka, Kosuke Kato, Naoki Nonaka ,Jun Seita. Efficient HLA imputation from sequential SNPs data by transformer. 2024 August 2.

[10] Sándor Baráth, Parvind Singh, Zsuzsanna Hevessy, Aniko Ujfalusi, Zoltán Mezei, Mária Balogh, Marianna Száraz Széles, János Kappelmayer. Enhancing HLA-B27 antigen detection: Leveraging machine learning algorithms for flow cytometric analysis. 2024 Feb 12.

[11] Seung-Hyun Jung, Sung-Min Cho, Seon-Hee Yim, So-Hee Kim, Hyeon-Chun Park, Mi-La Cho, Seung-Cheol Shim, Tae-Hwan Kim, Sung-Hwan Park, Yeun-Jun Chung. Developing a Risk-scoring Model for Ankylosing Spondylitis Based on a Combination of HLA-B27, Single-nucleotide Polymorphism, and Copy Number Variant Markers. 2016 Dec.

THANK YOU