Classification Of Autoimmune Disease Using Hybrid CNN-BiLstm Model

Ms.K.Kavitha[1], Aksharha S[2], Saraswathi K[3], Srinithi A[4], Naveena K[5]
[1]*Assistant Professor, Department of Computer Science and Engineering, Coimbatore Institute of Technology, Coimbatore*
[2,3,4,5]*UG Student, Department of Computer Science and Engineering, Coimbatore Institute of Technology, Coimbatore*

***Abstract*—Autoimmune diseases such as Type 1 Diabetes (T1D) and Multiple Sclerosis (MS) occur when the immune system mistakenly attacks the body's own tissues. Early detection of these diseases is important for effective treatment and disease management. T-cell receptor (TCR) sequences provide valuable information about immune responses and can be used to identify disease-specific immune patterns.**

This study proposes a computational framework for predicting autoimmune diseases using TCR sequence analysis and deep learning techniques. The dataset contains Complementarity Determining Region 3 (CDR3) sequences along with V-gene and J-gene information. Data preprocessing techniques are applied to clean and standardize the dataset. The biological sequences are then encoded using tokenization and padding techniques to convert them into numerical representations suitable for machine learning models.

A hybrid deep learning architecture combining Convolutional Neural Networks (CNN) and Bidirectional Long Short-Term Memory (BiLSTM) is used to extract sequence-level features and capture long-range dependencies in TCR sequences.

The system performs classification at the TCR sequence (cell) level, predicting whether each sequence belongs to Healthy, T1D, or MS classes. Final patient diagnosis is obtained by aggregating predictions from multiple sequences using majority voting. Experimental results demonstrate high classification accuracy, indicating the effectiveness of machine learning techniques for autoimmune disease prediction.

***Index Terms*—Autoimmune Disease Prediction, TCR Sequence Analysis, CNN-BiLSTM, Deep Learning, Bioinformatics.**

I. INTRODUCTION

Autoimmune diseases are conditions in which the immune system mistakenly attacks healthy cells, leading to serious health complications. Among the most common are Type 1 Diabetes (T1D) and Multiple Sclerosis (MS). Understanding and diagnosing these diseases remains challenging due to the complex nature of immune responses.

T-cell receptors (TCRs), particularly the Complementarity Determining Region 3 (CDR3), play a crucial role in antigen recognition and immune response. The high variability of CDR3 sequences makes them valuable for studying disease-specific patterns, but also difficult to analyze using traditional methods.

Conventional approaches often fail to capture the complex sequence patterns and dependencies present in TCR data, especially when datasets are limited and high-dimensional. To address these challenges, this work proposes a hybrid deep learning model that combines Convolutional Neural Networks (CNN), Bidirectional Long Short-Term Memory (BiLSTM), and an attention mechanism to effectively learn sequence features. Additionally, SHAP-based analysis is used to improve model interpretability by identifying the contribution of different features.

This integrated approach aims to enhance the accuracy and reliability of autoimmune disease classification using TCR sequence data.

II. LITERATURE REVIEW

1. Deep learning-based prediction of autoimmune diseases

Donghong Yang, Xin Peng, Senlin Zheng, and Shenglan Peng (2025) proposed a deep learning-based approach for predicting autoimmune diseases using TCR sequencing data. They utilized AutoY (CNN-based model) and LSTMY (BiLSTM with attention mechanism) along with multi-instance learning techniques. The study used the immuneACCESS dataset with top-100 abundant TCR sequences per patient. The model achieved outstanding performance with AUC values of 0.999 for Type 1 Diabetes and 0.996 for Multiple Sclerosis, demonstrating high effectiveness in distinguishing autoimmune patients from healthy individuals.

2. Autoimmune diseases: Current knowledge, immunological diagnosis and treatment

P. Mridula, P. Prasanth, P. Rahul, G. Saranya, M. Amrutha, and K. Rajalakshmi (2022) presented a study on autoimmune diseases focusing on immunological diagnosis and treatment. The work applied deep learning techniques such as YOLOv8 and Detectron2 for instance segmentation and classification of cell images. Using the ICPR 2016 Hep-2 Cell dataset, the study highlights the importance of automated cell analysis for accurate diagnosis and provides insights into immune system behavior in autoimmune conditions.

3. Prediction of Autoimmune Diseases Using Machine Learning Algorithms

Vandana, Amandeep Kaur, and Navdeep Kaur (2023) proposed a machine learning-based approach for predicting autoimmune diseases using clinical and immunological markers. The study implemented multiple algorithms including Support Vector Machine (SVM), Random Forest (RF), Logistic Regression (LR), and Naïve Bayes (NB). Among these, SVM and Random Forest showed superior performance compared to other models, demonstrating their effectiveness in disease prediction and classification tasks.

4. Artificial intelligence for autoimmune disease prediction: a systematic review

Chandra Prakash and Dinesh Kumar Vishwakarma (2025) conducted a systematic review on artificial intelligence for autoimmune disease prediction. The study analyzed various machine learning and deep learning techniques including Decision Trees (DT), SVM, Random Forest, CNN, RNN, and hybrid models. It highlights key challenges such as dataset limitations, class imbalance, and lack of standardization, while also identifying research gaps and future directions in AI-based autoimmune diagnostics.

5. T Cell repertoire diversity decreases in Type 1 Diabetes patients.

The 2016 paper "T Cell Repertoire Diversity Is Decreased in Type 1 Diabetes Patients" (Genomics, Proteomics & Bioinformatics) shows T1D patients have significantly fewer unique TCR clones and more highly-expanded clones (median 22 vs 1 in T2D for CD4 T cells, P=7.0E-08) than T2D patients and controls, analyzed via high-throughput sequencing of blood TCRβ CDR3 regions from 9 T1D, 4 T2D, and 6 healthy subjects. Shared HECs across T1D patients suggest common autoreactive clones, with potential to distinguish T1D from T2D though larger studies needed.

6. 2024 review shows TCR-Seq revolutionizes autoimmune disease research.

The *Journal of Autoimmunity* paper *"Bulk T cell repertoire sequencing (TCR-Seq) is a powerful technology for understanding inflammation-mediated diseases"* comprehensively reviews bulk TCR profiling methods (multiplex PCR, 5′RACE) and analysis pipelines (Shannon diversity, GLIPH2 clustering, incidence-based clonotype discovery) across MS, RA, IBD, demonstrating tissue-specific expansions and ML predictors for disease-associated TCR-pHLA interactions.

7. Machine Learning Approaches to TCR Repertoire Analysis

This paper reviews ML/DL methods for analyzing T cell receptor (TCR) repertoires from sequencing data to assess immune status, detect diseases (leukemia, CMV, COVID-19), and aid cancer vaccines. Key approaches include diversity metrics, sequence clustering (TCRdist/GLIPH), k-mer motifs with classifiers, generative models, and emerging deep learning (DeepTCR), addressing challenges like data sparsity and batch effects.

S. No	Title	Technique	Performance Metric
[1]	Deep learning-based prediction of autoimmune diseases	AutoY (CNN), LSTMY (BiLSTM + Attention), Multi-instance learning	AUC: 0.999 (T1D), 0.996 (MS)
[2]	Autoimmune diseases: Current knowledge, immunological diagnosis and treatment	YOLOv8, Detectron2	Dataset-based (ICPR 2016 Hep-2 Cell Dataset)
[3]	Prediction of Autoimmune Diseases Using Machine Learning Algorithms	SVM, RF, LR, NB	SVM & RF performed best
[4]	Artificial intelligence for autoimmune disease prediction: a systematic review	DT, SVM, RF, CNN, RNN, Hybrid Models	Review-based (No fixed metric)
[5]	T Cell Repertoire Diversity Is Decreased in Type 1 Diabetes Patients	TCRβ CDR3 from blood CD4/CD8 T cells (9 T1D, 4 T2D, 6 controls) → Illumina sequencing → Shannon entropy diversity + HEC (≥1%) analysis	T1D shows lower TCR diversity (CD4: P=7E-08 vs T2D/2.7E-09 vs controls; CD8: P=1.4E-04/7.6E-06) and more HECs (CD4 median 22 vs 1 in T2D).
[6]	Bulk T cell repertoire sequencing (TCR-Seq) is a powerful technology for understanding inflammation-mediated diseases	Bulk TCR-Seq (multiplex PCR/5′RACE from RNA/DNA) → NGS → Shannon diversity, GLIPH2/TCRdist clustering, Fisher exact test incidence analysis, DeepTCR/ML classifiers	DeepTCR AUC 0.85-0.90 (repertoire classification); pMTnet AUC 0.827 (ROC)/0.565 (PR) for TCR-pHLA binding.
[7]	Machine Learning Approaches to TCR Repertoire Analysis	Diversity metrics, TCRdist/GLIPH clustering, k-mer classifiers, generative models, DeepTCR for TCR repertoire analysis	general high ROC-AUC for CMV (~0.95+), autoimmune (~0.97), DeepTCR (~0.99).

III. DATASET

A. COLLECTION

The dataset used in this study consists of T-cell receptor (TCR) repertoire data obtained from three groups:

- Healthy individuals
- Type 1 Diabetes (T1D) patients
- Multiple Sclerosis (MS) patients

The dataset contains three CSV files:

- healthy.csv
- t1d.csv
- ms.csv

Each dataset contains information about TCR sequences including:

- CDR3 amino acid sequence
- V gene segment
- J gene segment

B. DATASET COMPONENTS

Each record in the dataset includes the following features:

FEATURE	DESCRIPTION
CDR3_SEQ	TCR amino acid sequence
V_GENE	Variable gene segment
J_GENE	Joining gene segment
LABEL	Disease class

Class labels are assigned as:

LABEL	CLASS
0	Healthy
1	T1D
2	MS

To ensure balanced training data, up to 10,000 sequences were selected from each class, resulting in a balanced dataset.

C. DAtA PREPROCESSING

Several preprocessing steps were applied to improve data quality:

- Selection of relevant features (CDR3 sequence, V gene, J gene)
- Removal of missing values
- Conversion of gene sequences to uppercase format
- Filtering of sequences with length less than six characters
- Standardization of dataset format

These preprocessing steps ensured consistency across the datasets and prepared the data for machine learning models.

D. FEATURE EXTRACTION

Key features were extracted from the dataset to represent biological sequence information:

CDR3 Sequence Features

CDR3 sequences were converted into numerical representations using character-level tokenization.

V Gene Encoding

V gene segments were encoded into numerical indices using tokenization.

J Gene Encoding

J gene segments were also encoded into numerical form using tokenization.

All sequences were padded to a fixed length of 26 characters to maintain consistent input dimensions.

The encoded features were stored as numerical arrays:

- X_seq.npy (CDR3 sequences)
- X_v.npy (V gene data)
- X_j.npy (J gene data)
- y.npy (class labels)

VI. SYSTEM ARCHITECTURE

A) Embedding Layer

The embedding layer converts input sequences (CDR3, V gene, J gene) into dense numerical vectors. This helps the model learn meaningful representations of biological sequences instead of using raw categorical values.

B) CNN (Convolutional Layer)

The CNN layer extracts local patterns (motifs) from CDR3 sequences. These motifs represent important biological patterns that are useful for distinguishing between disease classes.

C) BiLSTM (Bidirectional LSTM)

The BiLSTM layer captures long-range dependencies in the sequence by processing it in both forward and backward directions. This helps in understanding the contextual relationships between amino acids in the sequence.

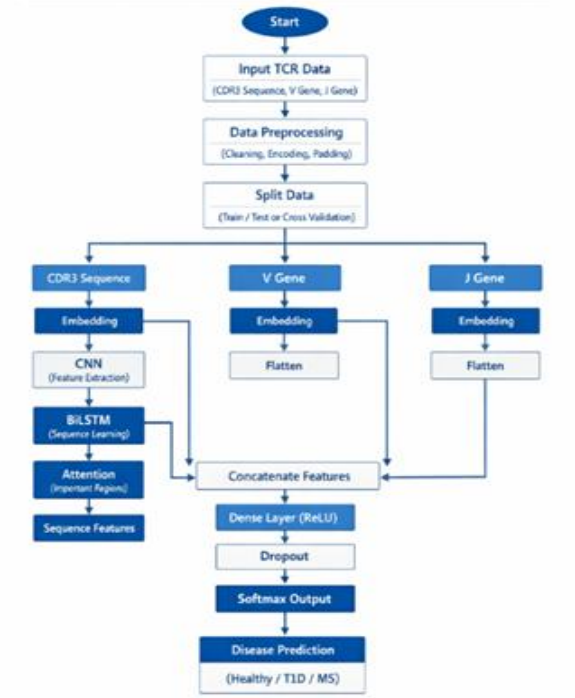


Fig 1: Proposed Hybrid CNN–BiLSTM with Attention Architecture for TCR-Based Autoimmune Disease Prediction

D) Attention Layer

The attention mechanism assigns importance weights to different parts of the sequence. It allows the model to focus more on biologically relevant regions, improving both performance and interpretability.

E) Feature Concatenation

Features extracted from the sequence branch and gene branches (V and J genes) are combined into a single feature vector. This integration improves classification by using both sequence and gene information.

F) Dense Layer (ReLU)

The dense layer learns complex patterns from the combined features and transforms them into a form suitable for classification.

G) Dropout Layer

Dropout randomly disables some neurons during training to prevent overfitting and improve model generalization.

H) Softmax Output Layer

The softmax layer converts the final output into probability scores for each class (Healthy, T1D, MS). The class with the highest probability is selected as the prediction.

V. IMPLEMENTATION

The system processes TCR sequence data to predict autoimmune diseases (Healthy, T1D, MS). Initially, datasets are cleaned, balanced, and preprocessed by removing missing values and filtering valid CDR3 sequences along with V and J genes.

Next, feature encoding is performed where CDR3 sequences are tokenized and padded, while V and J genes are converted into numerical form.

A hybrid CNN–BiLSTM with Attention model is used, where CNN extracts local patterns, BiLSTM captures sequence dependencies, and the attention layer highlights important sequence regions. V and J gene features are embedded, flattened, and combined with sequence features for final classification using a softmax layer.

Model performance is evaluated using 5-fold stratified cross-validation. Predictions are made at the sequence level and aggregated using majority voting for final patient diagnosis.

Finally, SHAP is applied to interpret the model by identifying important features influencing predictions.

BEGIN
INPUT:
TCR dataset (CDR3 sequences, V_gene, J_gene)

OUTPUT:
Predicted class (Healthy / T1D / MS)

STEP 1: Data Preprocessing
- Load dataset files (healthy.csv, t1d.csv, ms.csv)
- Select required columns:
- CDR3_seq
- V_gene
- J_gene
- Label
- Remove missing or invalid values
- Convert sequences and gene names to uppercase
- Filter sequences with length < 6
- Balance dataset by selecting equal samples from Healthy, T1D, and MS classes

STEP 2: Feature Encoding
- Apply character-level tokenization

→ Convert CDR3 sequences into integers
- Encode V_gene and J_gene using indexing
- Apply padding to CDR3 sequences

→ Fixed length = 26
- Generate feature arrays:

X_seq → encoded CDR3 sequences
X_v → encoded V gene
X_j → encoded J gene
y → class labels

STEP 3: Hybrid CNN–BiLSTM with Attention
SEQUENCE BRANCH:
- Input: X_seq
- Apply Embedding layer
- Apply CNN (Conv1D)

→Extract local motifs
- Apply Bidirectional LSTM

→Capture sequence dependencies
- Apply Attention Layer

→Assign importance weights
→Extract important sequence features
SET sequence_features = Attention output

GENE BRANCH:
- Input: X_v → Embedding → Flatten → V_features
- Input: X_j → Embedding → Flatten → J_features

FEATURE FUSION:
- Concatenate:

sequence_features + V_features + J_features

CLASSIFICATION:
- Apply Dense layer (ReLU)
- Apply Dropout (0.4)
- Apply Output layer (Softmax)

→ Classes: Healthy, T1D, MS

TRAINING:
- Loss Function = Categorical Cross Entropy
- Optimizer = Adam

- Metrics = Accuracy, Precision, Recall, F1-score

STEP 4: Model Evaluation
- Apply 5-fold Stratified Cross-Validation
- Evaluate performance on validation sets

STEP 5: Prediction
- Input encoded TCR sequence into model
- Compute class probabilities
- Select class with highest probability

STEP 6: Patient-Level Diagnosis
- Collect predictions for multiple sequences
- Apply Majority Voting

IF majority = Healthy
OUTPUT: Healthy

ELSE IF majority = T1D
OUTPUT: Type 1 Diabetes

ELSE
OUTPUT: Multiple Sclerosis

STEP 7: Model Interpretation (SHAP)
- Use SHAP to compute feature importance
- Analyze contribution of:

- Sequence positions
- V gene
- J gene

END

A. DEEP LEARNING MODEL

The proposed system leverages a hybrid deep learning architecture combining Convolutional Neural Networks (CNN) and Bidirectional Long Short-Term Memory (BiLSTM) for the prediction of autoimmune diseases such as Type 1 Diabetes (T1D) and Multiple Sclerosis (MS) using T-cell receptor (TCR) sequence data.

CNN is employed to automatically extract local sequence motifs from the encoded CDR3 amino acid sequences. The convolutional layers capture important short-range dependencies and biologically relevant patterns within the sequences, while pooling layers reduce dimensionality and retain the most significant features. This enables efficient representation of sequence-level information.

BiLSTM is integrated to capture long-range dependencies and contextual relationships within the TCR sequences. Unlike traditional models, BiLSTM processes the sequence in both forward and backward directions, allowing it to learn complex immunological patterns that may span across the sequence. This is particularly important for TCR data, where sequence context plays a crucial role in antigen recognition.

The hybrid CNN–BiLSTM architecture combines the strengths of both models—CNN for feature extraction and BiLSTM for sequential learning—resulting in improved classification performance. The final fully connected layers perform multi-class classification to predict whether the input belongs to Healthy, T1D, or MS categories.

Dropout regularization is applied to prevent overfitting, especially considering the limited dataset. The model is trained using the Adam optimizer and categorical cross-entropy loss function, ensuring efficient convergence and accurate prediction.

B. DEEP LEARNING MODEL

In the proposed system, the deep learning model is designed to perform multi-class classification of autoimmune diseases based on TCR repertoire data. The primary task is to classify input TCR sequences into one of three classes: Healthy, Type 1 Diabetes (T1D), or Multiple Sclerosis (MS).

The input to the model consists of encoded CDR3 sequences along with V-gene and J-gene features. These features are first preprocessed and transformed into numerical representations using tokenization and padding techniques to ensure uniform sequence length.

The CNN layers extract low-level sequence features such as motifs and conserved patterns, while the BiLSTM layers capture higher-level dependencies and sequence context. This combination allows the model to effectively learn both local and global representations of TCR sequences.

The final output layer uses Softmax activation to generate probability scores for each class. The class with the highest probability is selected as the predicted disease label.

Unlike image-based models, this system does not perform localization but focuses entirely on sequence-based classification. The integration of preprocessing, feature encoding, and hybrid deep learning significantly enhances the model's ability to identify

disease-specific immune signatures, making it suitable for computational immunology applications.

VI. FUTURE WORK AND LIMITATIONS

Despite achieving strong classification performance, the proposed model has certain limitations. The primary limitation is the small dataset size, with limited patient diversity, which may affect the model's generalization and increase the risk of overfitting.

Future work will focus on training the model with larger and more diverse TCR datasets to improve robustness and clinical reliability. Incorporating additional biological features, such as D gene segments and physicochemical properties of amino acids, may further enhance performance. Advanced architectures like transformer-based models can also be explored for better sequence representation.

Additionally, real-world clinical validation is necessary to evaluate the practical applicability of the model. Improving interpretability methods beyond SHAP can provide deeper biological insights and support medical decision-making.

SCREENSHOTS

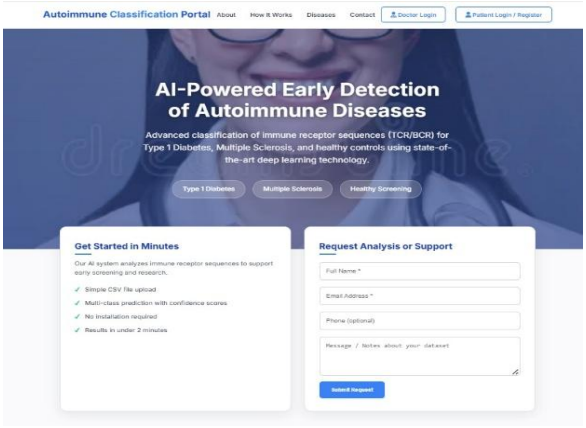
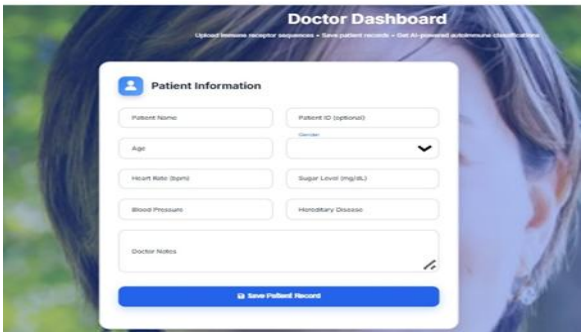


Fig 2: AI-Powered Early Detection Homepage



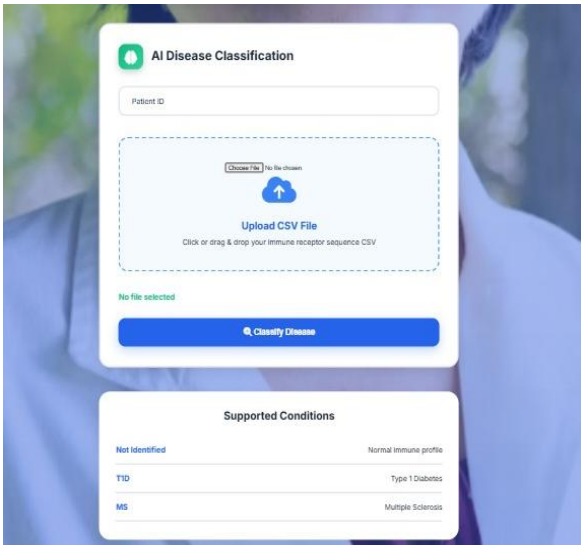


Fig 3: Doctor Dashboard Page

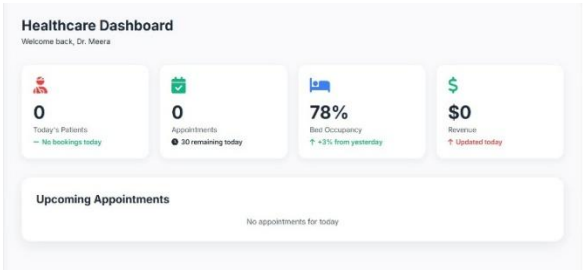


Fig 4: Healthcare Dashboard Page

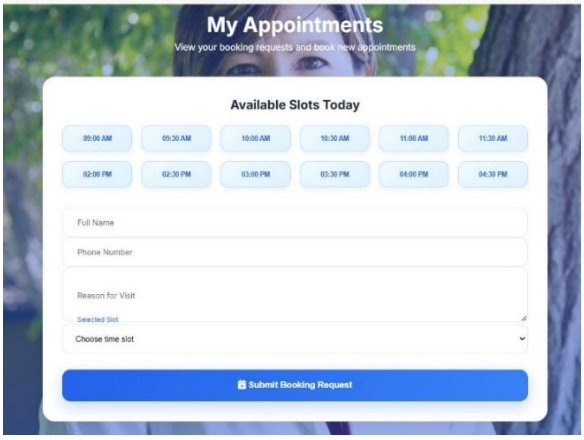


Fig 5: Appointment Booking Page

VII. RESULTS AND DISCUSSIONS

The hybrid CNN–BiLSTM model with attention achieved high performance in classifying TCR sequences into Healthy, T1D, and MS.

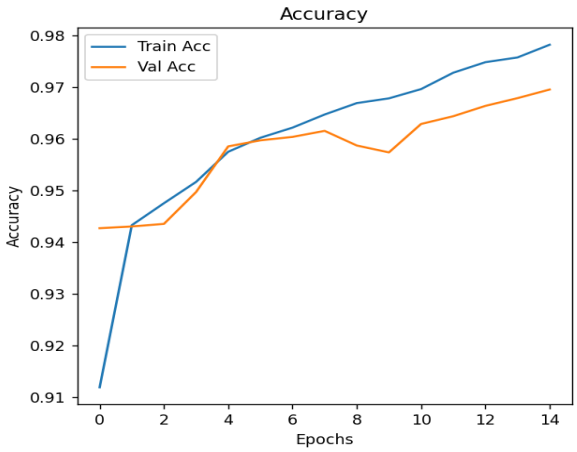


Fig 6: Training and Validation Accuracy over Epochs

The training and validation curves show stable learning, with accuracy reaching around 96–98% and loss decreasing consistently, indicating good model convergence with minimal overfitting.

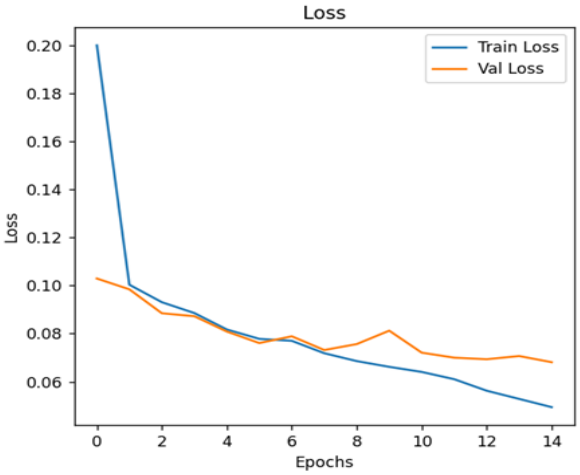


Fig 7: Training and Validation Loss over Epochs

The confusion matrix shows that most samples are correctly classified, with very few misclassifications and perfect prediction for the MS class.

The model successfully captures important sequence patterns using CNN and BiLSTM, while the attention mechanism improves focus on relevant features. SHAP further helps in understanding feature importance.

Overall, the model demonstrates strong performance even with limited data and provides a reliable approach for autoimmune disease prediction.

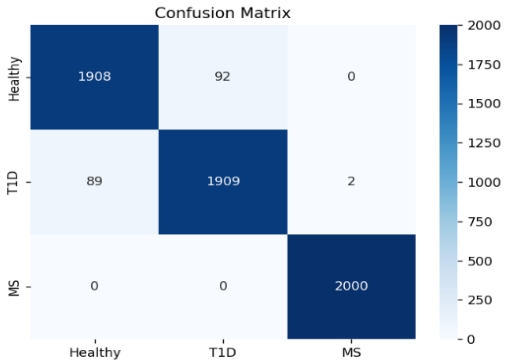


Fig 8: Confusion Matrix for TCR Classification (Healthy, T1D, MS)

REFERENCES:

[1] Yang, X. Peng, S. Zheng, and S. Peng, "Deep learning-based prediction of autoimmune diseases," Scientific Reports, Nature, 2025.

[2] B. Subramanya, D. B. Shivanna, N. Raj G, P. S. Prabhu, M. Yaseer, and R. S. Rao, "Detection of auto-immune disease using deep learning techniques," Mediterranean Journal of Rheumatology, 2025.

[3] D. Setiawan and R. N. Putri, "Prediction of autoimmune disease using backpropagation method," in Proc. CelSciTech-UMRI Conference, 2018.

[4] Vandana, A. Kaur, and N. Kaur, "Prediction of autoimmune diseases using machine learning algorithms," Lecture Notes in Networks and Systems, Springer, 2023.

[5] C. Prakash and D. K. Vishwakarma, "Artificial intelligence for autoimmune disease prediction: A systematic review," Artificial Intelligence Review, Springer, 2025.

[6] Y. Katayama, R. Yokota, T. Akiyama, and T. J. Kobayashi, "Machine learning approaches to TCR repertoire analysis," Frontiers in Immunology, vol. 13, p. 858057, 2022.

[7] A. K. H. Mahdy, E. Lokes, V. Schöpfel, V. Kriukova, O. V. Britanova, T. A. Steiert, A. Franke, and H. ElAbd, "Bulk T cell repertoire sequencing (TCR-Seq) is a powerful technology for understanding inflammation-mediated diseases," Journal of Autoimmunity, vol. 149, p. 103337, 2024.

[8] Y. Tong, Z. Li, H. Zhang, L. Xia, M. Zhang, Y. Xu, Z. Wang, M. W. Deem, X. Sun, and J. He, "T cell repertoire diversity is decreased in Type 1 diabetes patients," Genomics, Proteomics & Bioinformatics, vol. 14, no. 6, pp. 338–351, 2016.
[9] A. M. Mitchell and A. W. Michels, "T cell receptor sequencing in autoimmunity," J Life Sci (Westlake Village), vol. 2, no. 4, pp. 38–58, 2020.
[10] T. Shen, Y. Sheng, W. Nie, et al., "Deep learning-driven TCRβ repertoire analysis enhances diagnosis and enables mining of immunological biomarkers in systemic lupus erythematosus," BioData Mining, vol. 18, p. 76, 2025, doi: 10.1186/s13040-025-00490-5.