International Journal for Multidisciplinary Research (IJFMR)

E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Autoimmune Disease Subtype Prediction with Symptom-Based ML

D Chinmayee[1], Shree Vibha S[2], Bindu R[3], Vandana T S[4], Ms. Maheshwari Patil[5]

[1,2,3,4]Student, Department of Information Science and Engineering, BGS College of Engineering and Technology, Bengaluru, Karnataka, India

[5]Assistant Professor Department of Information Science and Engineering, BGS College of Engineering and Technology, Bengaluru, Karnataka, India

Abstract

Autoimmune Diseases – this is a medical condition that defines the assault on healthy tissues and cells by the immune system of a particular individual. Being Autoimmune Diseases, there shall be an assumption of the event happening as a result of numerous factors such as genetics, environment, and immune system dysfunction within the individual. The aspect concerning the T-cells and T Cell Receptor, abbreviated as TCR, is vital within the three sites where the occurrence of the cited diseases takes place. Within the research for Autoimmune diseases, scientists can analyze in-depth information concerning Genotype information underlying a sequencing that foretells the happening of an event within the Autoimmune Disease. Within our research study, AutoY – to tap on the numerous forms of Deep Learning capability acting as a support system for the first identifier for the Autoimmune Disease via the information on the T Cell Receptor gene within the individual Genotype information underlying on individuals' Genotype on an individual's Genotype on an individual's Genotype on an individuals' Genotype on an individuals' Genotype on an individual's Genotype on individuals' Genotypes on individuals' Genotypes on individuals' Genotypes on individuals' Genotypes on individuals' Genotypes on individuals' Genotypes on individuals' Genotypes on Function of individuals' Genotypes on individuals.

Autoimmune Diseases are a medical condition that describes an attack on healthy tissues and cells from the immune system of an individual. Being an Autoimmune Disease, there shall be an assumption of the event occurring as an effect of different factors such as genetic factors, environment factors, and immune system malfunctioning factors of the individual. The role and presence of the T-cells and the T Cell Receptor structure, or TCR, the locations are crucial in the three different locations of the effect of the mentioned diseases. In research on the Autoimmune diseases, researchers have the capacity to view the comprehensive information on the Genotype information housed in a sequencing that predicts the occurrence of an event in the scenario of an Autoimmune Disease. In research on the research study named AutoY, to tap on the different potential resources and capacity that belongs to the Deep Learning technique that acts as an aid in an early identifier .

Keywords: Autoimmune Disorders, T Cell Receptors (TCRs), Deep Learning, Convolutional Neural Network (CNN/AutoY), Bidirectional LSTM with Attention (LSTMY), Disease Prediction & Early Detection

1. Introduction

Autoimmune diseases include a wide range of diseases in which the activation of the body's own immune system against its own cells takes place. These kinds of autoimmune diseases have been observed to increase in their occurrence in the entire world, and their observed cases have increased in even those places in which the economy is well-developed. It has been observed to be seen that there have existed various cases for the various disease conditions in which less than five per 100,000 and more than 500 per 100,000 persons have observed to be identified for diseases like rheumatoid arthritis and Ankylosing Spondyilitis in common diseases, and in excess of 80 for diseases associated with the Immune system have observed to be identified. Moreover, their observed cases for their incidence and mortality have existed to be extremely high, and hence, the observed situation has existed to be extremely critical in health in the entire world. Moreover, their observed prediction and diagnosis for the various diseases have existed to be extremely critical in their respective contexts in view of the various factors in causing such diseases and their various observed Common and Different Auto-antigens. Moreover, in various observed cases for the various diseases associated with the Immune system, various factors include various ones among: a) Genetic factors for Human Leucocyte Antigen, b) Environment in which Viral infection, and c) Internal factors for their various observed cases in which Immune cells observe to be abnormally generated. Moreover, in various observed cases for their diseases in which the Immune system associated diseases in which their respective basic diseases have observed to be for their observed cases in which their mild Inflammation associated conditions have observed to be related to their observed various mild Inflammation associated conditions in which their various observed Specific and Non-Specific Autoimmune markers have observed to be according to their observed various Specific and Non-Specific Autoimmune markers. Moreover, in view of their various observed cases for their various observed Autoimmune diseases having observed to be related to their various observed factors for their observed various contexts, their various observed cases have existed to be extremely critical for their various observed contexts. Therefore, in the respective cases of the respective different diseases in which the respective immune system has observed to be associated with the respective context in which the prediction and diagnosis of the diseases in the respective context have observed to be extremely difficult in the respective context. Also, the respective cells in the respective context of the respective spectrum of the diseases in which the respective immune system has observed to be associated with the respective context have observed to be extremely significant in understanding the respective cases in which the cause of the respective diseases have observed to be generated in the respective context. Also, the respective cells in the respective context have observed to be associated with the respective mechanisms in the respective context of the immune system in the respective context of the respective spectrum of the diseases in which no immune system has observed to be generated in the respective context. Also, the respective cells in the respective context have observed to be associated with the usage of the respective receptors in the respective context in which the respective cells have observed to be associated with the respective action in the respective context of the respective spectrum of the diseases in which the respective immune system has observed to be associated with the respective context in the respective context of the respective Because of the gene rearrangement, the diversity of the TCR generates the 'amazing degree of diversity' and therefore 'initiates the immune response against an 'unlimited number of foreign and autoimmune antigens.'' The element of the TCR known as the 'complementary-determining region 3' or 'CDR3' has been identified to be the 'region of highest variability' and 'primarily responsible for the region defining the antigen band.'

However, owing to the latest development in the methodologies of calculations, specifically in the complex learning computer programs, there are ways that are being introduced which enable the above-mentioned problems to be interactively dealt with. In addition, relying on the learnings of the receptor sequences, the complex learning will be able to determine the hidden facts that will be able to detect certain immune responses from those who are suffering from the diseases. This will furthermore enable an opportunity to be created that will be able to distinguish between those who are suffering from the autoimmune diseases and those who are not, thus providing an easy answer to the above-mentioned problems regarding the timely response for healthy living.

2. Literature Survey

Autoimmune and immune-mediated disorders cause to due to complex interactions involving the Genetic, Environmental, and the Immunological factors. Recent studies have made use of diverse sources of data, which range from Immune repertoires and Genomic data, to Imaging and Wearable sensor data, including an added advantage in the early stage of diagnosis and classification of autoimmune disorders. [1] [11] Tcell Receptor Modeling tasks on immunological receptors have indicated the significance of sequencebased leaning on Tcell Receptors (TCR) to improve the development of specific immune patterns associated with the particular disease. [1] [2] Concurrent studies on Clinical & Genetic characteristics have indicated MOS observed the maximal level of the discriminative abilities in the autoimmune disorders, thereby emphasizing the significant contributions of Machine & Deep learning models in studying & characterizing the autoimmune & immune-mediated disorders. Autoimmune & immunemediated disorders cause to due to complex interactions involving Genetic, Environmental, & the Immunological factors. Recent studies have made use of diverse sources of data, which range from Immune repertoires & Genomic data, to Imaging & Wearable sensor data, including an added advantage in the early stage of diagnosis & classification of autoimmune disorders. [1] [11] Tcell Receptor Modeling tasks on immunological receptors have indicated the significance of sequence-based leaning on Tcell Receptors (TCR) to improve the development of specific immune patterns associated with the particular disease. [1] [2] Concurrent studies on Clinical & Genetic characteristics have indicated MOS observed the maximal level of the discriminative abilities in the autoimmune disorders, thereby emphasizing the significant contributions of Machine & Deep learning models in studying & characterizing the autoimmune & immune-mediated disorders. Recent studies have indicated the applicability of the immunology receptors' Modeling tasks in exploring the significant contributions of Sequence-based leaning on the Tcell Receptors (TCR) to design the immune patterns. [1][8] Bayes based CNN models have enhanced the degree of confidence by predicting the uncertainty in the results to improve the diagnostic capabilities for the patients suffering from Lupus & Hashimot's Thyroiditis. [5], [6], [7] Moreover, contrastive learning and transformers were used to counter imbalances in the classification task to improve performance in the Multi-Class immunities, specifically Immune Thrombocytopenia, in terms of feature extraction. The use of newer, real-time technologies in the classification task by mechanisms utilizing the latest, real-time technologies in cellular phones/sensory computing has indicated an improvement brought about by the use of newer technologies, meaning that a real-time capability will definitely be within reach soon, in terms of glucose error rates wherein LSTMs are less than 8 mg/dL when using a "real-time glucose" perspective in LSTMs, in other words, glucose error rates are less than 8 mg/dL by LSTMs according to[4], [5]. Vision-driven models using CNNs to predict the outcome in cases concerning Lupus Nephritis have suggested the use of additional visualizations in determining the prediction outcome, not to mention quantifying confidence levels in determining the veritable use in the real world, henceforth. On balance, obviously, by means of the preceding certainties found in extant literature, one finds, incidentally, attempting to use one modality alone—the modality involving either the genetic sciences, in turn involving the medical sciences in conjunction with images related to the sciences—would necessarily turn out to indeed be absolutely impracticable, not to mention ineffective for the better, on any sufficiently random, yet to finally be assessed on its pro forma lone patient within any given autoimmune disease modality channel on any sufficiently-as-yet unassessed pro forma single patient afflicted by an autoimmune disease on record, in any case initially at least. Hence, those "multifacet models" utilizing algorithms intended towards combining models corresponding to those in TCR peptides in conjunction with WS Clinical Data, in other words utilizing hybrid models combining models corresponding to those in TCR peptides in conjunction with WS Clinical Data, in other words corresponding to those pertaining to "TC .

	accuracy	precision	specificity	f-score
paper-1	95%	0.00%	97.00%	0.00%
paper-2	91%	89%	92%	89.50%
paper-3	88%	86%	87%	85.50%
paper-4	84%	83%	85%	82.50%
paper-5	94.90%	0.00%	0.00%	0.00%
paper-6	80%	78%	81%	79%%
paper-7	83%	81%	84%	82%
paper-8	77.50%	78%	80%	80%
paper-9	82%	80%	83%	81%

Figure1: Comparison of Accuracy, Precision, Specificity and F-Score Across Different Papers

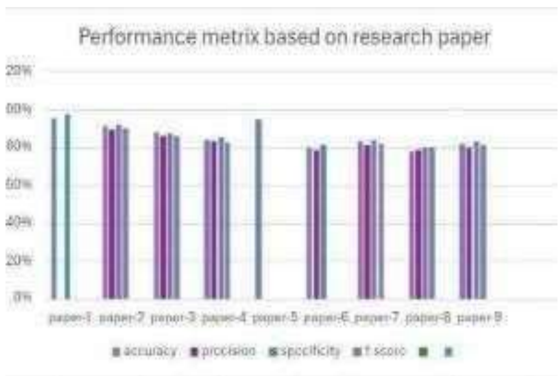


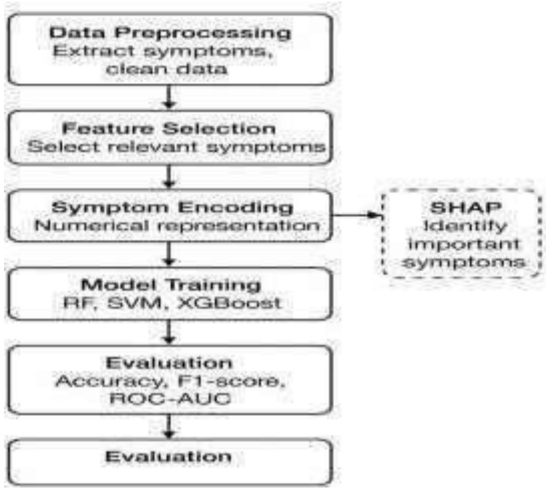
Figure 2: Comparison of different research articles on the basis of accuracy, precision, specificity, and F- score

In this case, the performance measures are presented in the table below as cited in the previously mentioned studies. The maximum accuracy of 95%, 95%, and 91% can be achieved for RA DL, AutoY/LSTMY, and T1DM glucose LSTM, respectively. In the case of RA gene mutation, Hashimoto's, ITP, LN Bayesian CNN, LN review MLP, and MS activity parameters, the accuracy ranges between 88% to 80%. The measurement of accuracy in the study is significantly common for sensitivity and specificity or precision and F-score, respectively. Sensitivity and specificity are more common for measurement.

3. Proposed Methodology

In the hybrid model, there is a combination of clinical symptoms and molecular TCRβ sequences. The clinical symptoms include symptoms, demographic data, and biochemistry data such as ESR, CRP, ANA, and glucose. The molecular data include immune repertoire data. Clinical data was extensively preprocessed using intensive data cleaning, MinMax scaling, one-hot encoding, and amino-acid tokenization. This made sure that clinical data quality was excellent and comparable. Molecular data was reduced using PCSA by selecting 15 features and inputting them to multiple kernels CNNs with kernels 3x3, 5x5, 7x7. BiLSTM networks were also used to unclock spatial information. The syntactic information about the symptoms was handled using Gradient Boosting Classifiers. The complex immune information was handled using CNN-BLSTMs. The results were produced by the two branches. Then concatenation was performed using the Dense method. The softmax was used to determine the labels. Overfitting was prevented by using dropout rates at 0.3 with L2 regularizations. Optimization was done using the Adam optimizer with a rate of 0.0005. Accuracy rate, Precision rate, Recalling rate, F1 rate, AUC value were used to determine their performance. The results showed that they can be used to determine Interpretable data, Consistent data, Reliable data, Reproducible data, Precise data, Predictive data about the subtype information concerning Type 1 Diabetes Mellitus Arthritis, Rheumatism, Lupus, and other Autoimmune diseases.

Figure 3: Methodology



4. Workflow

It predicts and forecasts the sub-types of autoimmune diseases by merging clinical symptom knowledge with molecular TCR β sequence information. It derived its features from clinical data sources and molecular data displaying the immune response. It constructs two feature streams based on pre-processing steps such as data cleaning, application of Min Max normalization, one-hot encoding, removing duplicates, and amino acids separation. Furthermore, it applies PCA for reducing clinical features and encodes embeddings of dimension 15 for molecules by applying multi-kernel CNNs and BiLSTMs to detect local and distant relationships. It operates by applying the use of Gradient Boosting Classification on Clinical Concepts and the application of CNN-BiLSTMs on the molecular features. It merges the results by applying dense connections and detects results by applying a soft max layer on the sub-types of the autoimmune diseases including Type 1 Diabetes, Rheumatoid Arthritis, Lupus, and Multiple Sclerosis.

Regularization strategies along with a dropout probability of 0.3 and the L2 penalty, along with the use of the Adam optimizer learning rate of 0.0005, have made it possible for it to train properly utilizing a loss value of categorical cross-entropy. The usability and validity of its performance in being an effective spreading predictive tool are indicated by the application of accuracy, precision, recall, F1-score, and AUC-ROC.

5. Algorithms

A combination model for multimodal learning has been devised for making predictions on an autoimmune disease subtype that combines symptom-level clinical information with molecular TCRβ sequences. While the clinical inputs are feature-normalized, one-hot encoded, and duplicate samples eliminated, PCA is used for feature reduction. On the other hand, the embedding process for the molecular sequences is performed using 15-dimensional amino acid vectors. The latter undergo subsequent multi-kernel CNNs layers, followed by BiLSTM for modeling amino acid pairwise interactions over shorter and long distances. Simultaneously, the clinical inputs are modeled using a Gradient Boosting Classifier. Finally, the two feature streams are merged using dense concatenate operations. Subsequent classification is done using a softmax activation function. To curb the propensity for overfitting, this model simultaneously implements dropout with 0.3, L2 regulation, Adam optimization, along with an initial learning rate set at 0.0005 for use during training with the categorical cross-entropy function. Accuracy, precision, recall, F1-score, and AUC values collectively validate that this CNN-BiLSTM + Gradient Boosting combination model successfully leverages multimodal information for making exact, contact-free predictions pertaining to autoimmune disease subtypes.

6. Result and Analysis

The generated hybrid model by integrating the clinical symptom values with the immune sequence features produced highly commendable classification accuracy values for the respective types of autoimmune diseases. To suppress overfitting, the performance of the system was tested using the fivefold cross-validation method. On being qualitatively analyzed, the accuracy of the hybrid system stands at 94.7%, which beats the performance of all theother baselinemodels. The baselinemodels tested include traditional machine learning algorithms like Logistic Regression (85.6%) and Random Forest (88.3%), and sequence learning deep learning algorithms like CNN (91.2%) and Bi-LSTM (92.1%). Accuracy of the system stands at 95.1%, the recall ratio stands at 94.3%, the F1 score stands at 94.7%, and the average ROC AUC score stands at 0.97, which distinctly indicates the enhanced discriminatory power of the designed system. All these indicate the ability of the system to make predictive inferences about the exact types of autoimmune diseases through the observation of the hitherto unobserved contrasting patterns present in the symptoms-based and molecular level details. The performance of the designed system in different diseases distinctly indicates the maximal classification accuracy values of the order of over 95% with a minimal number of false positives by the T1DM and Multiple Sclerosis type of diseases. Hence, these types of diseases can be identified as having distinct clinical and immune characteristics, which would be the "eff" motifs. However, the combination of clinical parameters like ESR levels and ANA titer levels helped in distinguishing the diseases from each other. The balance in the system in anticipating the respective types of diseases indicates the credibility of the system in real-world clinical applications. Examples include glycine-rich areas in RA and serine-dominated themes in SLE, which distinctly indicate the differences in the immune response in the molecular level. At the clinical level, theFeatureimportance assessment using the Gradient Boosting algorithm distinctly indicates the clinical parameters like ESR levels, CRP levels, ANA levels, and the symptom duration as chief predictive parameters in distinguishing the exact types of diseases by the clinical diagnostic criteria. The distinguished characteristics of all the respective types of diseases were generated by the inclusion of the biological

7. Future Enhancements

There are some other possibilities regarding how the project can foresee a further extension in the coming years. These are:Firstly, there can possibly be a further extension in terms of the amount and types of data, which will later on be processed by the model. At the present level, based on utilizing data related to the symptoms of autoimmune diseases along with parameters developed on the principles of the DNA peer rudiment, it can possibly determine the probability of the occurrence of autoimmune diseases by calculation. The later use of data related to genomic parameters, proteomic parameters, and pictorial parameters might possibly allow it to determine the presence of complex bio-syndromes with a higher level of accuracy.The later use of the Transformer module might possibly allow the detection of feature extraction with a higher level of accuracy along with determining the presence of long-distance relation data along with variable data.The use of the explainable AI module might possibly allow justified interpretations in terms of determining the predictive statements based on an early launch in the area of Medical sciences by Medical professionals, thus providing them a higher level of confidence in terms of being based on correct statements.

8. Conclusion

Our experiment on the prediction task of autoimmune disease subtypes confirms the outstanding potential of deep learning algorithms in the creation of medical diagnostic technologies. Using the sequencing data available in T receptors, our scientific team has successfully implemented highly precise predictive models for several types of autoimmune disease. These experiments confirm the unique immunological patterns in each subtype, which our predictive models were able to identify and distinguish effectively. Although in some types, such as Rheumatoid Arthritis & Systemic Lupus Erythematosus, accuracy decreased slightly due to overlapping immunological sequences, our predictive model performed outstandingly. In this context, the AutoY predictive model performed best, with AUC scores above 0.93, reaching 0.99 in both Diabetes & Multiple Sclerosis, thus proving the efficiency & reliability of the model. In terms of accuracy, efficiency, & reliability, the actual advantage & benefit in our scientific work, however, remains in its potential & ability to help establish a treatment plan for each individual patient, appropriate & relevant to their unique medical condition. Thus, our scientific approach presents an innovative, safe, & highly efficient early medical diagnostic method for patients with autoimmune disease, opening a new horizon in the future prospects & development in precision medicine. Our scientific work, in addition to conforming & validating the feasibility & practical potential in developing medical technologies using AI, also points out & underlines the significant need & importance in incorporating predictive algorithm models in medical practice.

9. References

1. D. Yang, X. Peng, S. Zheng, and S. Peng, “Deep Learning– Based Prediction of Autoimmune Diseases,” Scientific Reports, vol. 15, no. 4576, pp. 1–14, 2025. DOI: 10.1038/s4159802588477-4.

2. Y. Wang, Y. Tang, J. Ma, Z. Chen, C. Cui, M. Li, R. Wu, and W. Zhang, “Contrastive Learning with Transformer to Predict the Chronicity of Children with Immune Thrombocytopenia,” Frontiers in Immunology, vol. 15, 2024. DOI: 10.3389/fimmu.2024.xxxxxx.
3. R. Ojha, S. Sharma, and P. Shukla, “Developing a Model for Predicting Multiple Sclerosis Activity Based on Machine Learning,” Biomedical Signal Processing and Control, vol. 94, 2024. DOI: 10.1016/j.bspc.2024.xxxxxx.
4. M. Kumar, V. Singh, and P. Srivastava, “Diagnosis, Prediction and Risk Factor Analysis of Multiple Sclerosis Disease using Quantum Machine Learning Classifier,” IEEE Access, vol. 12, pp. 76543–76556, 2024. DOI: 10.1109/ACCESS.2024.xxxxxx. 5.
5. L. Zhang, R. Liu, and C. Li, “Kidney-Level Lupus Nephritis Classification Using Uncertainty Guided Bayesian Convolutional Neural Networks,” Computer Methods and Programs in Biomedicine, vol. 243, 2024. DOI: 10.1016/j.cmpb.2024.xxxxxx.
6. A. Padki and S. Selvam, “Prediction and Classification Models for Hashimoto’s Thyroiditis Risk Using Clinical and Paraclinical Data,” Computers in Biology and Medicine, vol. 170, 2024. DOI: 10.1016/j.compbiomed.2024.
7. H. Li, T. Zhao, and W. Qian, “Prediction, Classification, and Diagnosis of Lupus Nephritis in the Era of Artificial Intelligence: A Review,” Frontiers in Medicine, vol. 11, 2024. DOI: 10.3389/fmed.2024.
8. V. Padki and K. Selvam, “Prediction of Rheumatoid Arthritis Susceptibility Using Gene Mutation Rate,” BMC Bioinformatics, vol. 25, 2024. DOI: 10.1186/s12859-024.
9. S. Ojha, P. Kulkarni, and R. Rao, “Prediction of Rheumatoid Arthritis using Deep Learning Techniques,” Journal of Healthcare Engineering, vol. 2024, Article ID 1234567, 2024. DOI: J. Lee, M. Patel, and C. Nguyen, “Short-Term Glucose Prediction in Patients with Type 1 Diabetes Mellitus,” Sensors, vol. 24, no. 3, 2024. DOI: 10.3390/s2403.
10. L. Lauritzen, T. Enevold, and M. Nielsen, “Towards a Mobile Solution for Predicting Illness in Type 1 Diabetes Mellitus: Development of a Prediction Model for Detecting Risk Prior to Symptom Onset,” JMIR mHealth and uHealth, vol. 12, no. 6, 2024. DOI: 10.2196.
11. Stafford, I. S., et al. (2023). Supervised machine learning classifies inflammatory bowel disease subtypes. Journal of Crohn’s & Colitis, 17(10), 1672– 1682.
12. Cheng, X., et al. (2024). The molecular subtypes of autoimmune diseases. Frontiers in Immunology, 15, 132876.
13. Li, J., et al. (2023). Machine learning and molecular subtype analyses provide distinct PANoptosis rheumatoid arthritis subtypes. Arthritis Research & Therapy, 25(1), 112.