International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Comparative Analysis of Machine Learning Algorithms Techniques for Human Immunity Level Prediction

Anugraha Manoj[1], Remya Ranganath[2]

[1,2]Department of Mathematics, Amrita School of Physical Sciences, Amrita Vishwa Vidyapeetham, Kochi

Abstract:
The early detection of disease can be crucial as well as challenging. Due to immune diseases our immune system get damaged and it cause harmful to our body. For predicting human immunity levels, the study brings up the combination of ML techniques with the help of the data set with 20,853 entries (rows) and 33 columns (features). For bridging AI and Healthcare together we use Supervised and Unsupervised Learning techniques. Also this study applies Cross-Validation, Regularization. pruning, and early stopping to ensure that the models don't overfit, which is the key to challenge towards AI and Healthcare. Among these Linear Regression was the best model for this specific dataset and could predict immunity levels accurately . The usage of MSE tells us how much error this model makes in predicting immunity levels. R^2 tells about the model fitting helps in making useful predictions.

Keywords: Machine Learning, Autoimmune,Healthcare,Immunity, Mathematical Modelling, cross validation

INTRODUCTION

The human immune system is highly complex and dynamic defence mechanism whereby it fights harmful pathogens, such as bacteria, viruses, and others, inside the body to keep them out from the body. It is a well-coordinated network of cells, tissues, and signaling molecules that distinguish between what constitutes self and non-self elements. However, there are instances in which it malfunctions, resulting in autoimmune diseases. These are conditions in which the immune system mistakes the body's own cells for attacks themselves, leading to chronic inflammation and tissue damage. (KELLERMANN, et al., 2020)[1]

Autoimmune diseases are highly heterogeneous. They can target different organs and systems of the body. Some, like type 1 diabetes, are focused on a particular organ, whereas others, such as systemic lupus erythematosus (SLE), affect several tissues at once. The most common autoimmune disorders are rheumatoid arthritis, multiplesclerosis, Hashimoto's thyroiditis, Crohn's disease, and psoriasis. The specific causes are unknown, but findings indicate that a genetic susceptibility along with factors such as environmental exposures, infections, hormonal imbalances, and gastrointestinal tract microbial disruptions contribute to the development and progression of these diseases. Women predominate, whose immune systems may be affected by sex hormones. (Male)[3]

The generality of autoimmune diseases has increased globally and imposed a considerable burden on

health systems. Late diagnosis and absence of curative treatment further worsen the condition of the disease and often results in long-term disability and a diminished quality of life. Present methods of diagnosis use biomarkers, imaging, and clinical symptoms. (goodnow)[2] However, this may not result in early detection. Hence, what is now greatly needed are far more precise predictive measures for a healthier immune system to determine persons at risk while they are symptom-free.

The developments of artificial intelligence and machine learning have opened new possibilities for understanding immune system behaviour and predicting immune-related disorders. (KELLERMANN, et al., 2020)[1] (goodnow)[2] By analyzing large datasets, including genetic profiles,biochemical markers, lifestyle patterns, and environmental exposures, machine learning algorithms can identify patterns that may not be easily detected through traditional methods. These predictive models have the potential to revolutionize personalized medicine, enabling early intervention, tailored treatment strategies, and better disease prevention.

This research will take a look at the potential application of machine learning models in predicting human immunity levels. Here, we are trying to develop a models that can show the immunity level as a function of multiple physiological and environmental parameters using advanced computational techniques. The results of the study lead towards more efficient healthcare solutions concerning patient's outcomes and well-being.

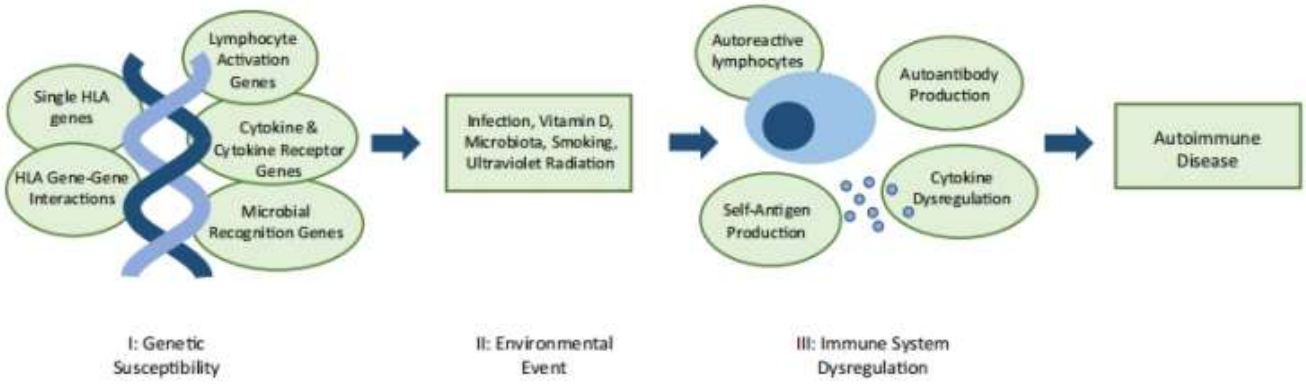


Fig 1: Pathogenesis of Autoimmune Diseases: Genetic, Environmental, and Immune Factors. (KELLERMANN, et al., 2020)[1]

SUPERVISED AND UNSUPERVISED LEARNING

ML is a subfield of AI that equips computers to learn patterns from data and make decisions with minimal human intervention. ML algorithms can be broadly classified into Supervised and Unsupervised learning based on the nature of the training data and the learning process.

Supervised Learning

It means a model gets trained on labelled data, which involves each input sample having its known output. In this approach, the algorithm learns by making an input mapping into an output through patterns discovered within the data. The prime goal is the reduction of differences between predicted values and actual ones; techniques may involve regression or classification. Some common algorithms for supervised learning include LinearRegression, DecisionTrees, SupportVector Machines (SVM), Neural Networks. It includes

1. **Healthcare:** Disease prediction based on patient data.
2. **Finance:** Credit risk assessment and fraud detection.
3. **Natural Language Processing (NLP):** Sentiment analysis and language translation.

Unsupervised Learning:

Here the data does not have output labels. That is there is no specific output that needs to be given. The algorithm finds the hidden-patterns, structures, or relationships in the data without any kind of guidance. This is highly useful for exploratory data analysis and tasks that require hidden structures to be unearthed. It includes

1. **Customer Segmentation**: It involves grouping similar customers based on purchasing behaviour.
2. **Anomaly Detection:** It detects fraudulent transactions or network intrusions.
3. **Recommendation Systems:** Personalize content suggestions in streaming platforms.

Both supervised and unsupervised learning are vital for modern AI-driven applications. Supervised learning is ideal for predictive modelling tasks, while unsupervised learning is beneficial in finding hidden insights within large datasets. The choice between these methods depends on the problem at hand, data availability, and desired outcomes.

Avoiding Overfitting:

Machine learning overfitting is a common occurrence where a model learns noisy or random fluctuation in training data rather than the underlying pattern. It leads to high training data accuracy but low generalization to new, unseen data. Overfiting is usually overfiting when a model is extremely complex relative to the datasets , producing memoirs rather than learning.

Techniques to avoid overfiting:

Many strategies can be employed to reduce overfitng and improve the normalization capacity of the model:

1. **Cross validation:**

The K-folded cross-validation helps to assess the model performance in many of the most of the dataset , ensuring that it is not highly dependent on specific data points.

2. **Regularization:**

Techniques such as L1(Lasso)andL2(Ridge) add penalty conditions to the loss function, preventing extreme complexity by shrinking the model weight.

3. **Pruning (decision of trees):**

Decision reducing the depth of trees helps eliminate unnecessary branches that can catch noise instead of useful patterns.

4. **Feature Selection and Engineering:**

Removing irrelevant or fruitless features reduces model complexity and reduces the risk of overfitng.

5. **Dropout(for nerve network):**

The dropout disables a portion of neurons during training forcing the network to learn more strong and generalized representatives.

6. **Stop early:**

Monitor the performance of the model on verification data and stop performing when the performance deteriorates from fitting to noise.

Large datasets provide more representaive patterns,making the model missing the model.

Data growth technique can also help in cases where it is challenging to get more data.

7. **Learning to dress:**

Many models (eg: increase and boost techniques such as random forest and gradient boosting) increases

generalization by reducing dependence on predictions of any single model.

METHODOLOGY

This study gives various **machine learning techniques** to analyze and predict outcomes based on input data. The methodology is structured into **data preprocessing, model selection, training, evaluation, and visualization**, using appropriate computational tools.

1. **Tools and Technologies Used To implement and evaluate the models, the following tools were utilized:**

- **Python**: The primary programming language used for model development.
- **Scikit-learn**: A machine learning library for implementing regression and classification models.
- **NumPy & Pandas**: These are the libraries used for numerical computations
- **Matplotlib & Seaborn**: Library that used for analyzing model performance graphically.
- **Google Colab**: Used for executing and testing the code.

2. **Data Preprocessing**

For generalising the model, the data is divided into training and testing subsets. Then feature training and normalization like preprocessing techniques are applied to improve the model performance.

3. **Model Selection and Implementation**

The machine learning models like Linear regression, KNN, Decision Tree, and Random Forest are used for comparison. Each model was implemented using Scikit-learn, ensuring consistency in algorithm execution and performance evaluation.

4. **Model Training and Evaluation**

The models were trained using the training dataset, and their performance was assessed using standard evaluation metrics – MSE and R-squared score for measuring the quality of predictor and evaluates how good the model explains the variance in the data.

5. **Visualization and Comparative Analysis**

Using Matplotlib and Seaborn, The graphical interpretation of the result is done.

Discussion:

The evaluation and comparison of the predictive performance of multiple machine learning models with Linear Regression, KNN, Decision Tree, Random Forest, SVR, and K-Means Clustering are done. The models were trained and tested using structured numerical data, and their effectiveness was measured using Mean Squared Error (MSE) and R-squared (R^2) Score.

1. **Model Performance Comparison**

The result indicates that linear regression performed well rather than the others with relatively low error and a high R^2 score(See Table 1). Being an unsupervised learning model, K-Means Clustering helped in identifying patterns within the dataset.

2. **Implications of Findings**

The findings suggest that the choice of regression model depends on the nature of the dataset. As the data follows a linear pattern, Linear Regression is sufficient, rather than Random Forest and SVR. The study also gives the importance of model evaluation using multiple measures to ensure the correct decision-making in predictive analytics. A comparative visualization of the predicted and actual levels of immunity showed that Linear Regression closely followed true values, further validating its accuracy.

The experiment proves that machine learning models can accurately predict human levels of immunity. Linear Regression worked best because it could perfectly model the dataset, thus being extremely appropriate for these conditions.

3. Limitations and Future Work

1. The data used was small in size; a larger dataset would further confirm the findings.
2. Hyperparameter tuning and deep learning methods can be investigated for better accuracy.
3. Future research can incorporate real-world clinical data to improve model robustness.

RESULTS

Graphical representation:

Table 1: Performance of various machine learning models was assessed by Mean Squared Error (MSE) and R-squared Score (R^2). Lower MSE reveals less error, and higher R^2 reflects good predictability.

Model	Mean Squared Error (MSE)	R-squared Score (R^2)	Remarks
Linear Regression	0.0000	1.000	Perfect prediction; potential overfitting
XGBoost Regression	0.1859	0.976	Excellent performance, realistic accuracy
K-Nearest Neighbors	~4.5	Negative	Poor prediction, lacks generalization
Decision Tree Regression	~2.75	Negative	Overfits on training data
Random Forest Regression	~3.63	Negative	Moderate overfitting, unstable results
Support Vector Regression	~11.41	Negative	Weakest performance overall

1. Linear Regression

Linear Regression explores the relationship of independent variables to the dependent variable (immunity level) that is linear in nature. It forecasts immunity using a straight line that results in the minimal gap between estimated and observed values.

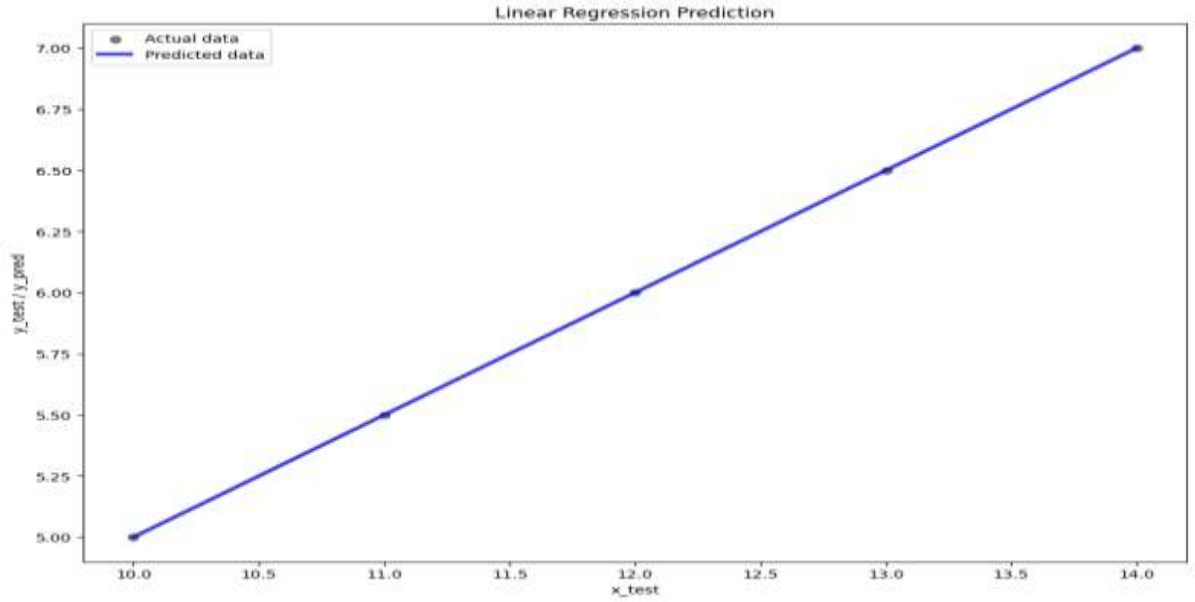


Fig 2: Actual vs Predicted Immunity Scores using Linear Regression

2. K-NEAREST NEIGHBORS(KNN)

KNN forecasting immunity involves finding 'k' data points (neighbors) with known immunity levels nearest to the subject in question. The predicted level of immunity is based on the average immunity of the neighbors.

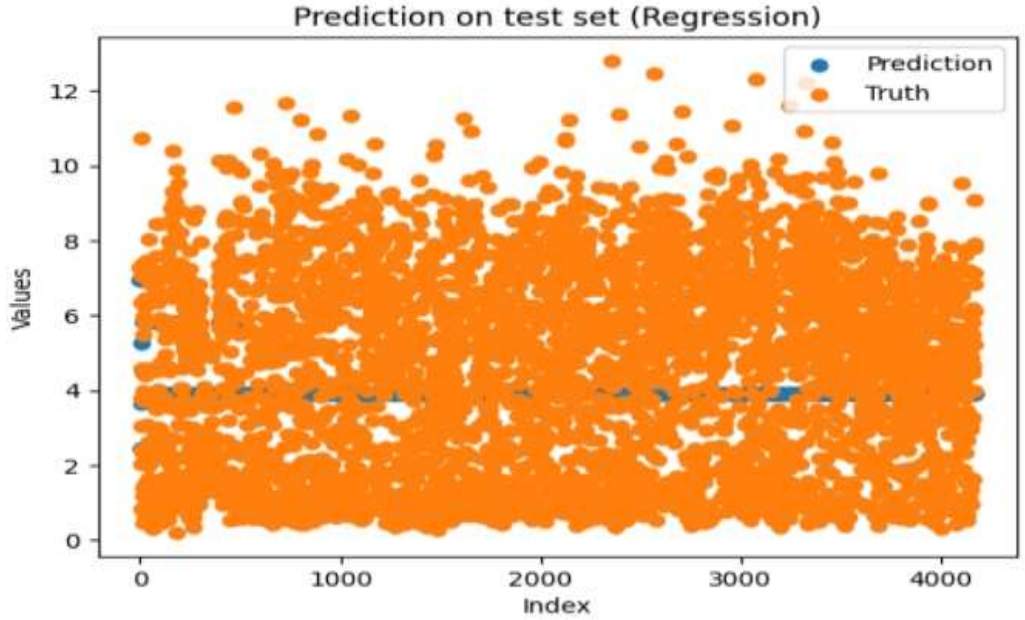


Fig 3: Shows the predicted vs actual immunity scores using KNN .the predicted value is blue and are nearly flat and do not show the variation in the true value (orange) this shows KNN has failed to get the meaningful understanding from the data which resulted in poor predictive accuracy

3. Decision tree

Performance:

MSE = ~2.75

R^2 = Negative

4. Random Forest Regression

Performance:

MSE = ~3.63

R^2 = Negative

5. Support Vector Machine (SVM)

Performance:

MSE = ~11.41

R^2 = Negative

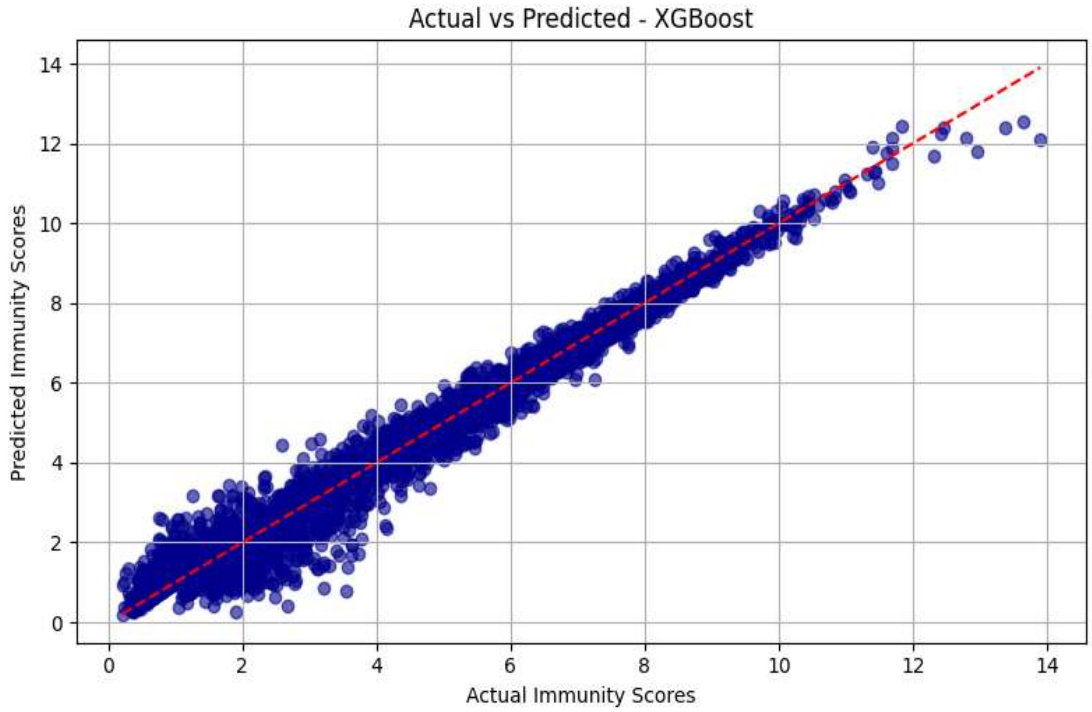
6. XGBoost

It's a machine learning technique that helps in building multiple decision tress and combines their output for more accurate predictions.

Performance:

MSE = 0.1859

R^2 = 0.9764



Key Findings

1. Linear Regression obtained the minimum MSE (0.0) and maximum R^2 (1.0), making it the best fit model for this dataset.
2. The KNN, Decision Tree, and Random Forest models gave the negative values of R^2, means that their generalization is very poor.
3. K-Means Clustering was utilized for unsupervised learning and does not offer classical MSE/R^2 but may assist in identifying patterns.

CONCLUSION

After going through the models using the metrics MSE and R^2 the followig were intepreted :

1. Linear Regression showed perfect results on the dataset (R^2 = 1.0, MSE = 0.0) indicating a strong linear pattern.However the potential of being overfitted cannot be overlooked.
2. XGBoost showed as the most reliable model with R^2 of 0.9764 and MSE of 0.1859 .The result demonstrated that XGBoost was able toaccurately tell the model relationship between the input features and immunity levels while maintaining the robustness .
3. Other models like KNN ,SVM and Random Forest showed poor performance with negative R^2 values and significantly larger error rates indicating lack of fit with this dataset.
4. In Conclusion the results validate the applicability of Machine Learning particularly XGBoost in health prediction and immunological analysis.The study shows that the computaional models can give accurate, and efficient tools for immunity evaluation which helps in early detection of the disease.

REFERENCES

1. Stafford, I. S., Kellermann, M., Mossotto, E., Beattie, R. M., MacArthur, B. D., & Ennis, S. (2020). A systematic review of the applications of artificial intelligence and machine learning in autoimmune diseases. NPJ digital medicine, 3(1), 30.
2. Goodnow, C. C., Sprent, J., de St Groth, B. F., &Vinuesa, C. G. (2005). Cellular and genetic mechanisms of self tolerance and autoimmunity. Nature, 435(7042), 590-597.
3. Kuchroo, V. K., Ohashi, P. S., Sartor, R. B., &Vinuesa, C. G. (2012). Dysregulation of immune homeostasis in autoimmune diseases. Nature medicine, 18(1), 42-47.
4. Cooper, G. S., Bynum, M. L., & Somers, E. C. (2009). Recent insights in the epidemiology of autoimmune diseases: improved prevalence estimates and understanding of clustering of diseases. Journal of autoimmunity, 33(3-4), 197-207.
5. Hayter, S. M., & Cook, M. C. (2012). Updated assessment of the prevalence, spectrum and case definition of autoimmune disease. Autoimmunity reviews, 11(10), 754-765.
6. Eaton, W. W., Rose, N. R., Kalaydjian, A., Pedersen, M. G., & Mortensen, P. B. (2007). Epidemiology of autoimmune diseases in Denmark. Journal of autoimmunity, 29(1), 1-9.
7. Cho, J. H., & Feldman, M. (2015). Heterogeneity of autoimmune diseases: pathophysiologic insights from genetics and implications for new therapies. Nature medicine, 21(7), 730-738.
8. Simon, T. A., Kawabata, H., Ray, N., Baheti, A., Suissa, S., & Esdaile, J. M. (2017). Prevalence of co-existing autoimmune disease in rheumatoid arthritis: a cross-sectional study. Advances in therapy, 34, 2481-2490.
9. Gilhus, N. E., Nacu, A., Andersen, J. B., & Owe, J. F. (2015). Myasthenia gravis and risks for comorbidity. European journal of neurology, 22(1), 17-23.
10. Ruggeri, R. M., Trimarchi, F., Giuffrida, G., Certo, R., Cama, E., Campennì, A., ... &Wasniewska, M. (2017). Autoimmune comorbidities in Hashimoto's thyroiditis: different patterns of association in adulthood and childhood/adolescence. European journal of endocrinology, 176(2), 133-141.
11. Fatima, M., & Pasha, M. (2017). Survey of machine learning algorithms for disease diagnostic. Journal of Intelligent Learning Systems and Applications, 9(01), 1.
12. Dalgleish, A. (1999). The relevance of non-linear mathematics (chaos theory) to the treatment of cancer, the role of the immune response and the potential for vaccines. Qjm, 92(6), 347-359.

13. Ma, Y., Chen, J., Wang, T., Zhang, L., Xu, X., Qiu, Y., ... & Huang, W. (2022). Accurate machine learning model to diagnose chronic autoimmune diseases utilizing information from B cells and monocytes. Frontiers in immunology, 13, 870531.