Next-Generation Vaccines: Leveraging Deep Learning for Predictive Immune Response and Optimal Vaccine Design

[1]Saranya K R, [2] Josephine Usha L, [3]Valarmathi P and [4]Suganya Y
[1,2,4]Department of Computer Science and Engineering, School of Computing, SRM Institute of Science and Technology
Tiruchirappalli, Tamil Nadu, India.
[3]Department of Computer Science and Engineering, Mookambigai College of Engineering,
Pudukkottai, Tamil Nadu, India.
[1]dr.k.r.saranya@gmail.com, [2]josephineusha@gmail.com, [3]vgoodmathi@gmail.com, [4]suganyasuchithrra@gmail.com

Corresponding should be addressed to Saranya K R : dr.k.r.saranya@gmail.com

Article Info
Journal of Machine and Computing (https://anapub.co.ke/journals/jmc/jmc.html)
Doi: https://doi.org/10.53759/7669/jmc202505061
Received 12 April 2024; Revised from 23 November 2024; Accepted 23 January 2025.
Available online 05 April 2025.

Abstract – The rapid advancement in vaccine development has become increasingly critical in addressing global health challenges, particularly in the wake of emerging infectious diseases. Traditional methods of vaccine design, while effective, often involve lengthy processes of trial and error, which can delay the deployment of life-saving immunizations. In the pursuit of enhancing vaccine efficacy, the application of deep learning techniques has emerged as a transformative approach. This study presents the development and implementation of an Integrated Neural Network Model (INNM), which synergistically combines Artificial Neural Networks (ANNs) and Random Forests for predictive immune response and optimal vaccine design. The INNM employs a hybrid feature selection methodology, integrating Pearson correlation with Recursive Feature Elimination (RFE), to identify the most relevant immunological predictors. Implemented in a Jupyter Notebook environment, the model achieved an impressive accuracy rate of 98.4%, demonstrating its potential to revolutionize vaccine development. This innovative approach underscores the capability of deep learning to predict immune responses with high precision, paving the way for the next generation of vaccines.

Keywords – Deep Learning, Predictive Immune Response, Optimal Vaccine Design, Integrated Neural Network Model, INNM, Artificial Neural Networks, ANNs, Random Forests, Hybrid Feature Selection, Pearson Correlation, Recursive Feature Elimination.

I. INTRODUCTION

Predictive immune response and optimal vaccine design are a revolutionary concept in the setting of immunology and vaccine discovery. Through the use of sophisticated computational analytics and biological knowledge, this field can be used to predict what immune responses may arise from different antigens, informing how future vaccines should be developed in order to generate the most specific/high-affinity effects [1] [2]. Traditional vaccine development, however is largely empiric in nature and involves much trial and error. Predictive modelling can access an enormous data jungle of immunological sources (genetic, proteomic and epidemiological) to simulate what happens when the immune system reacts [3]. This in turn allows researchers to identify candidate antigens more quickly, and to design vaccine candidates with improved accuracy.

At its core, this relies on appreciating the diversity of immune systems. Innate and adaptive immunity are the two major divisions of immune response to pathogens, HLA (human leukocyte antigen) diversity being a decisive factor [4]. Incorporating genetic profiles into predictive models are able to consider this variation, can shape the development of broadly effective vaccines which work across different populations and genetic backgrounds. These models can also be used to predict side effects that the vaccines might cause, and thus improve their safety. Predictive immune response and the design of an ideal vaccine are far-reaching implications. This method supports rapid vaccine design and implementation to address global pandemics, such as COVID-19 [5] [6]. It also shows great potential for addressing longstanding problems like HIV and malaria, diseases that have stymied conventional vaccine approaches. The promise of a new field at the intersection among bioinformatics, systems biology and immunology is up to revolutionize vaccine development opening opportunities for more effective as well as personalized vaccination [7].

The best vaccination strategy and optimal delivery mechanisms including adjuvants to boost immune response also should be better elucidated. By developing computational tools that predict the interactions of formulations with immune cells, researchers are provided detailed guidance on which combinations will be most effective. These models also help to predict not only the length of the immune response but its strength, assisting in creating long-lasting vaccines [8] [9]. Designing an optimal vaccine is a complex and intricate endeavour that combines different scientific fields to deliver not only efficacious, but also safe and affordable vaccines. The key to this process is recognizing the most suitable antigens - substances that evoke an immune response [10] [11]. Current developments in genomics and proteomics are making it possible for researchers to have access to pathogens from the smallest of molecular levels, identifying specific antigen targets which most probably trigger a strong and protective immune reaction [12] [13]. The comprehensive insight into the biology of pathogens assists in choosing antigens that can propagate immunity, rather than pathology.

Adjuvants, which are substances that can contribute to activating the immune system of a host responding to an antigenic challenge, form part of the essential components in designing more effective vaccines. Choosing the right adjuvant (or collection of them) can increase a vaccine's effectiveness manifold because such adjuvants boost immune response so that we need far less antigen. This means that vaccines are not only cheaper, they have fewer associated side effects [14] [15]. Adjuvant aims to identify compounds that can specifically stimulate the immunological pathways of interest and therefore direct an immune response for a more efficient clearance of targeted pathogens. The delivery system for a vaccine is another important component of its design. Vaccines traditionally have been administered by injection; however, new technologies are emerging to consider other forms of vaccines such as nasal sprays, oral tablets and microneedle patches [16]. The need for ease of administration, improved patient compliance and potent mucosal immunity - critical to pathogens that enter the body via a mucous membrane - led us to investigate alternate pathways. New delivery technologies could also help vaccines remain shelf-stable, with knock-on effects for their accessibility to populations in low-resource settings [17] [18].

The idea of tailored vaccines is making headlines now. Customized vaccines based on specific genetic and immunological profiles of an individual, population group or even region could be developed to generate the most ideal response. The latter is especially attractive in the context of diseases such as cancer because tumour-specific antigens can be manipulated and thus a very individual, highly effective treatment can be performed [19]. Optimal vaccine design also requires thorough testing and validation. These are then followed by a number of phases of clinical trials — including in vitro experiments and animal model testing to determine the vaccine's safety, immunogenicity consistency/efficacy. Increasingly, computational modelling and simulations are used to predict outcomes and improve vaccine candidates in advance of clinical testing-speeding efforts while saving lives. **Fig 1** shows the benefits.

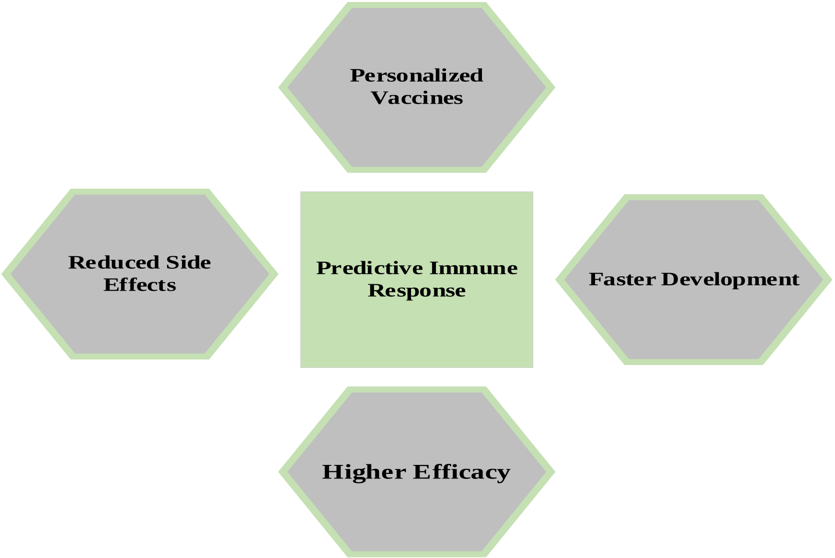


Fig 1. Benefits of Predictive Immune Response**.**

Predictive immune response is a forefront field in immunology that aims to utilize computational tools and biological data to predict how the immune system will reply when facing different types of antigens. This approach goes beyond traditional empirical methods based on trial and error using large datasets of genetic, proteomic, immunological data. By better modelling the complex dynamics of immune defences, scientists can predict more accurately how effective vaccines and therapies will be. Such predictive abilities are even more critical for practice in emerging infectious diseases and personalized medicine that require prompt/ accurate responses to identify a majority of effective treatments as well as prevention options. Using computational algorithms, like predictive modelling that examines extensive immunological data-such as genetic and proteomic information or epidemiology-a profile of the immune responses can be developed to give us a bird's eye view. This gives researchers an edge to be able to locate antigen targets with potential and engineer vaccine candidates more accurately-faster. It is an elaborate process that entails various advantageous scientific tools, new strategies and artistic sciences to fetch in being a highly effective vaccine which should be low cost as well. This process leverages genomics, adjuvant research, innovative delivery systems and personalized medicine for the development of safe vaccines that can generate powerful immune responses to protect against multiple diseases across all ages - significantly enhancing public health globally. This paper explores the integration of deep learning techniques, specifically Artificial Neural Networks (ANNs) and Random Forests, to predict immune responses and optimize vaccine design. The Integrated Neural Network Model (INNM) proposed in this study leverages these powerful computational tools to analyze complex biological data, identify key immunological features, and predict the efficacy of vaccine candidates. A significant aspect of the INNM is its hybrid feature selection methodology, which combines Pearson correlation and Recursive Feature Elimination (RFE). This approach ensures the selection of the most relevant features, enhancing the model's predictive accuracy.

Main Contribution of the Work

The primary contribution of this work lies in the development and validation of the Integrated Neural Network Model (INNM), a novel approach that combines Artificial Neural Networks (ANNs) and Random Forests to predict immune responses and optimize vaccine design. The key contributions are outlined as follows:

- The INNM synergistically integrates ANNs and Random Forests, leveraging the strengths of both models to enhance predictive accuracy. This hybrid approach capitalizes on the ANN's ability to model complex, non-linear relationships and the Random Forest's robustness against overfitting and high-dimensional data.
- The study introduces a hybrid feature selection process that combines Pearson correlation with Recursive Feature Elimination (RFE). This dual-step approach ensures the identification and retention of the most relevant immunological features, thereby improving the model's performance and interpretability.
- The INNM was implemented in a Jupyter Notebook environment, which supports reproducibility and accessibility. By providing a detailed and transparent implementation, this work enables other researchers to replicate the study and apply the INNM to various vaccine development contexts.
- By harnessing the predictive power of deep learning, the INNM offers a robust framework for accelerating vaccine development. This approach facilitates the rapid identification of promising vaccine candidates and the prediction of their efficacy, potentially leading to faster and more effective responses to emerging infectious diseases.

This paper is structured as follows: Section 2 reviews related work in vaccine design and deep learning applications in immunology. Section 3 details the design and implementation of the Integrated Neural Network Model (INNM), including the hybrid feature selection methodology. Section 4 presents the experimental setup and results, highlighting the framework's effectiveness in predicting immune responses and optimizing vaccine design. Finally, Section 5 concludes the paper with a discussion of future research directions and potential improvements to the INNM framework.

II. RELATED WORK

The implications for vaccine design are huge that can predict whether a peptide will be presented on MHC class I molecules. There is already a lot of very accurate peptide presentation predictions for MHC class I molecules that are based on deep learning. As they are black-box functions, very little is being known about the decision-making of these MHC class I predictors. To trust these forecasters requires not only an understanding of their rationale but also the ability to explain in a way humans can understand. AneXplainable AI (XAI) methods is implemented to help interpret MHC class I predictor outputs in the context of input peptide features. They offered experimental data that explains the results presented by four leading MHC class I predictors on a large dataset of MHC alleles and peptides. In addition, they evaluate the credibility of these explanations by comparing them with observed data and testing their robustness. MHCXAI seeks to improve this confidence by offering the most sophisticated machine leaning-based predictions through validated interpretations and enriched knowledge in immune response domain.

TCR sequencing has recently been used to profile the immune response or immunity towards cancer. Regrettably, most of the other research focused on quantitative indicators such as clonality and have largely ignored the complementarity-determining region 3 (CDR3) sequence. A deep learning system of algorithm, DeepTCR, to find sequence that predict response to immunotherapy. They demonstrated that DeepTCRwhich is capable of predicting the response of a patient and use the model to infer antigenic specificities forming part of the predictive signature and how they evolve during therapy. Non-responses have a greater diversity in their tumour-specific TCRs over the course of treatment compared with responders, whereby a high proportion of expected antitumor antigen recognizing TCrs response prediction signature. Their findings support a biological concept that accumulation of tumour-specific T cells undergoing treatment-mediated turnover is associated with nonresponse, potentially due to the defective state of these t-cells.

The most common way by which hepatitis E virus (HEV) is transmitted, an RNA-virus, thus leading to fecal contamination of water. The disease is considered to rank as second only headache to be the largest current public health risk in the world, especially low resource worlds; Africa being one of most affected countries. An African vaccine was expected to be essential for preventing infection with HEV. In-silico epitope based subunit vaccination is employed with CTL, HTL and BL epitopes fused to adjuvants/linkers. The vaccine candidate designed in silico had acceptable solubility and physiochemical properties, was found to be immunogenic while being non-allergenic as well as showing no signs of toxicity. Results showed stable binding efficacy, and MD simulation indicated the same interaction preservation. The vaccination will induce human immunological responses - as was inferred from immune simulation. These were subsequently integrated in silico into pET28 b (+) cloning vector Validation studies using in vitro and in vivo methods would be required to confirm this conclusion, but altogether these data strongly argue for the potential of an epitope-based subunit vaccine as prophylaxis against HEV.

With the advent of next-generation sequencing, it has become relatively straightforward to detect somatic mutations and develop patient-specific neoantigen cancer vaccines targeting unique tumor variations. These vaccines have the ability to produce meaningful therapeutic responses since they boost our immune system and help it fight off cancer cells. It has also proved difficult to determine the appropriate dose of vaccine specific for each patient because tumours come in so many flavours. To address this challenge, a mathematical model is used which describes the immune response cascade in an individual due to vaccination and formulate a dosage optimization problem on top of this. Theyoffered a protocol for the optimization of dosing strategies such that across immunizations, they minimize total tumour burden and activated T cells relative to an alternative repeated-dosing scenario. To validate their approaches, they performed in silico trials on six real patients with advanced melanoma from clinical trials. They examined the results of an appropriate dose of vaccine and how they compare to a suboptimal one. By tweaking the vaccination schedule with higher start doses and lower final doses to minimize bystander activation, optimal control of tumour growth may be more readily achieved in some patients using our models.

The virus that causes COVID-19 is the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2). Viruses like SARS-CoV-2 mutate all the time. The cost vaccine offers little or no protection against the omicron SARS-CoV-2 variant because of just how much its spike protein is mutated. Most vaccines against SARS-CoV-2 are also dependent on wild-type virus spike protein sequences. This increases the risk of a shift in the virus, making booster shots ineffective. Ultimately, the research will lead to a predictive vaccine and epitope discovery that guide reverse-translational modifications of the current sequences for vaccines. In this regard, epitopes derived from the spike proteins of wild-type, delta variant and omicron variants most probable as those are already major circulating or potential new combination containing any one/link between each other combinations to be emerged within these layers were employed in designing predictive vaccine by taking immune informatic approach. The vaccine was safe and induced an immune response. The vaccine antigen has been injected for 1 month, results of the C-ImmSim simulation indicate there is a sufficient level of humoral response and cell-mediated responses. The results suggest that the vaccine was effective and provided a sufficient level of immunity, the study says. It is suitable for the creation of antibodies or other forms, and can then be tested experimentally to develop a vaccine.

III. METHODOLOGY

Several key steps were taken during the development of Integrated Neural Network Model (INNM). Comprehensive data on immunologic correlates and vaccine response outcomes were first gathered. Dataset pre-processed - normalization, missing values treated We therefore used a hybrid feature selection approach that combined Pearson correlation for finding linear relationships with Recursive Feature Elimination (RFE) to eliminate progressively less important features. The INNM combined Artificial Neural Networks (ANNs) with Random Forests. This is done because ANNs are able to model nonlinear complex relationships as well, and the key advantage that Random Forests offer which helps them deal with overfitting due to high-dimensional data. Model training and validation was performed using stratified k-fold cross-validation to ensure generalizability. **Fig 2** shows the architecture of proposed model.

Dataset: Immune Response Dataset (IRD)

Immune Response Dataset (IRD), this dataset provides detailed immune response results from 86,723 individuals as well as the vaccine administration data. The data supports an array of factors that can impact and be used to evaluate vaccines as well the ensuing immune responses. Details of vaccines, HLA types and demographic information such as age/gender are part of the dataset that lists each entry. The dataset also documents the quantity of antibodies discovered, and how badly subjects experienced any side effects. IRD phases have been rigorously worked through with a combination of clinical trials, public health records and laboratory data. The data themselves undergo exhaustive preprocessing in order to ensure they are clean and correct. The preprocessing consisted of dealing with missing values, scaling numerical data and encoding categorical variable. Therefore, the dataset establishes a robust baseline for understanding how different variables influence vaccine-induced immune responses

Immune responses differ in various population segments, with demographic characteristics such as age and gender playing a crucial role. Addition of HLA types contributes towards genetic diversity and aids in understanding personalized vaccine responses. Information such as the type of vaccine, antigen and adjuvant is crucial to predicting how distinct components in different vaccines formulations influence immune function. The immune response, measured by antibody levels that indicate how well the vaccine works, is the primary outcome variable. The degree of these side-effects is also captured in order to evaluate vaccine safety.

Data Preprocessing

We conducted preprocessing steps to be able to analyze and model the Immune Response Dataset (IRD) as follows:

Data Cleaning

The data cleaning process was started by handling missing values which is a very important step to take as it ensures that our dataset remains reliable and we are able to further use this dataset. Some common possibilities that could lead to missing values are: data entry errors, loss during the transfer of information from one system or before compilation, unrecorded responses for personal reasons etc. For numeric variables, missing values were filled with the median imputation in the IRD Median was selected since it is less influenced by outliers and skewed data as compared to mean, therefore furnishes a more sturdy measure of central tendency. The missing values in the categorical variables by using mode. By doing this the categorical data remained close to most of the other points in terms their actual form, preserving some semblance of distribution and structure. What imputation did was to package both the size of dataset and prevented further analyses or machine learning models from having gaps in their data. Median Imputation of Missing Value for Numerical Columns:

$$x_{imputed} = Median(x) \tag{1}$$

Mode Imputation for Categorical Variables:

$$C_{imputed} = Mode(C) \tag{2}$$

Data Standardization and Normalization

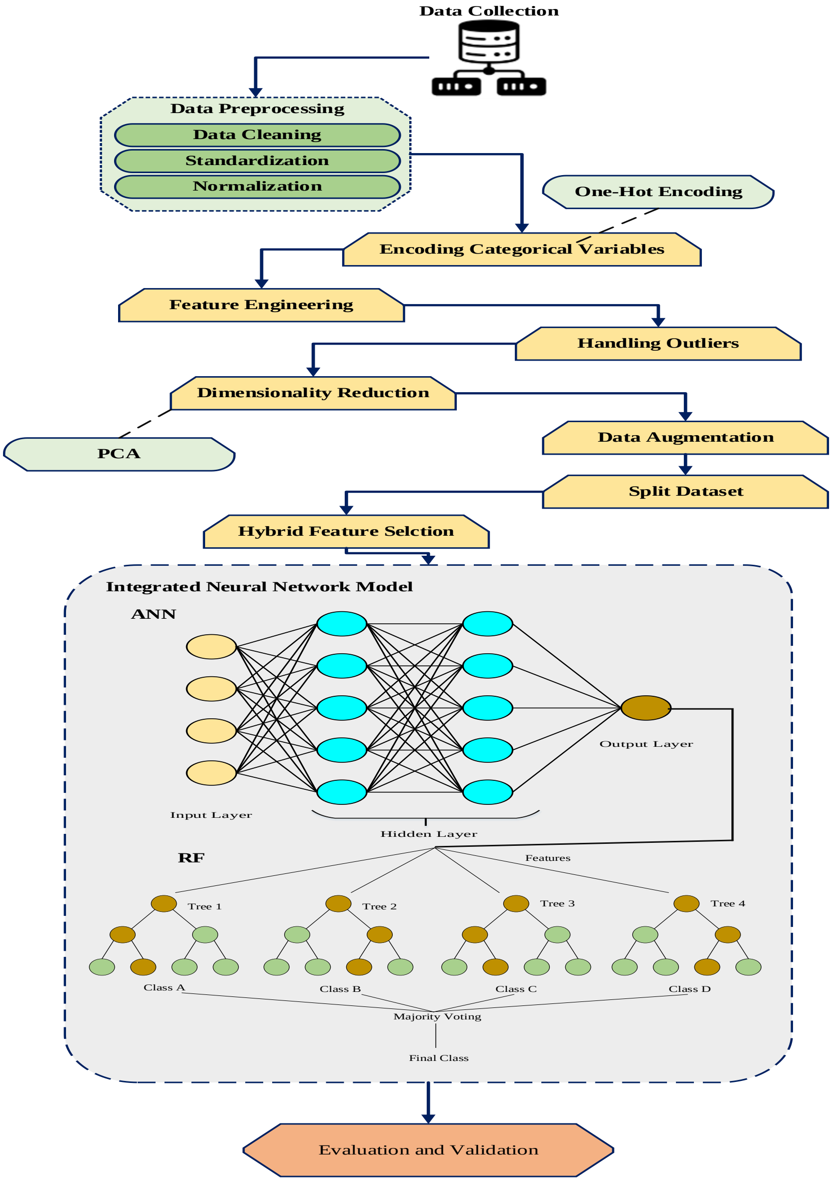


Fig 2. Architecture of Proposed Model.

Z-score normalization was carried out to standardize the numerical features for analysis. This normalization changes the numerical data such as age, and antibody levels to have a mean of zero and Standard Deviation 1. Normalization of numerical features is important as it allows every feature to be equally significant in the analysis, not letting those with larger scales have an undue favoring over them. In datasets where features differ considerably in scale (e.g., from age measured via years vs. antibodies measure with arbitrary units), models like k-nearest neighbors or neural networks may be biased towards larger-scale characteristics. Z-score standardization was applied to the numerical data in order that they were all on comparable scale, and then fed into machine learning algorithms for improved accuracy. This step also helped to have gradient-decent based optimization algorithms converge faster and more reliably, helping the overall training process. Z-Score Normalization:

$$z_i = \frac{x_i - \mu}{\sigma} \tag{3}$$

Encoding Categorical Variables

The immunological Response Database (IRD) contains the categorical variables like gender, HLA type or vaccine type; these are then one-hot encoded. It created binary vectors for each category within a feature where they can be converted to format which is accessible by those algorithms that are not able process categorical data on their own. The best example of that was one-hot encoding, because it kept the model from ascribing ordinal relationships to the categories and there is no such thing between those labels in this case. The HLA types or vaccine categories, as an illustration, are unique classes not having any implied set. Because each category was represented as a separate binary variable (0 or 1), the model could learn from all these features without making any incorrect assumptions about their relationships. This increased the dimensions of dataset because it converted all categorical data into One Hot encoded instead but left us with manageable dimensionality. This step becomes crucial for algorithms such as linear regression, support vector machines that expect numerical number. Binary vector representation:

$$v_j = \begin{cases} 1 & if\ the\ feature\ value\ is\ j \\ 0 & otherwise \end{cases} \tag{4}$$

Feature Engineering

Interaction terms were created to represent the relationships of features with other features. The intersection terms are the new inputs which forms by combining two or more existing input variables and learn from those things done nothing but correlation among features. If appropriate interaction terms were plausible (for instance vaccine type with adjuvant, or HLA type with antigen), these two way interactions might be included in the IRD as well. That could yield clues about how various blends of vaccine constituents and genetic factors could impact the immune response. Some HLA types might interact with an antigen in such a way that the immune response would be either more robust or much less strong and this is crucial for predicting which vaccines will work. These interaction terms were multiplied or added together stat-wise and the resulting features went back in to the data set. This step added more features to the dataset that might now benefit in providing a better prediction.

Handling Outliers

The dataset was also check for outliers using the Interquartile Range(IQR) method which is one way to detect extreme values in data that differs greatly from most of it The IQR method consists of calculating a band where 50% of the points (from median) from there forth, as outliers if such data lie below Q1-1.5*IQR or above Q3 + 1.5*IQR Once the outliers had been identified, they were clipped to a maximum threshold so as their presence does not over power the analysis. This capping method was useful for keeping the distortion that extreme values could cause especially in statistical analyses and machine learning models. This way the dataset was able to keep being robust, which guaranteed that models trained on this data would be more dependable and less impacted by outliers since we capped them out.

Balancing the Dataset

This step involved checking to see if there was any class imbalance in the dataset, mainly regarding categorical variables (e.g., our target or severity of side effects). An imbalanced dataset might cause the model to be biased and perform well on majority class while it works poorly near the minority. SMOTE (Synthetic Minority Over-sampling Technique) was used to handle this. SMOTE constructs examples on existing minority class examples by interpolating the feature space. It does so by balancing the dataset in a way that is more general than simply duplicating instances of the minority class which could result in overfitting. SMOTE generated synthetic data points to have fair representation of the minority classes so that model was able to learn from all sides. This was an important step to improve the performance of our classification algorithms and handle predictions in a fair and accurate manner across classes. Synthetic data generation:

$$x_{synthetic} = x_i + \lambda \cdot (x_{nn} - x_i) \tag{5}$$

Dimensionality Reduction

PCA (Principal Component Analysis) was used in this case to reduce the number of dimensions such that most of the variance is preserved. PCA is used to transform original features into a set of new orthogonal components by ordering them from the one which explains more variance in data. This reduction in dimensionality simplified the models, and decreased computational complexity by virtue of gleaning information from only identifying principal components that account for most of the variance. In the meantime, this also served to scale more data and reduce multi-collinearity among features. This dataset in particular was suited well to PCA, as it significantly improved the interpretability and functionality of basic machine learning models by removing some redundant or uninformative features. Projection onto Principal Components

$$X_{reduced} = XV_k \tag{6}$$

Data Augmentation

We generated the synthetic data for under-represented class using different kind of data augmentation techniques, in particular SMOTE. This step was important since the initial dataset is unbalanced and we need to make sure that enough examples are provided for all classes so our model can be trained it properly. Generating synthetic data tackled the extremely low examples of certain combinations and diversified training set to expose model more scenarios. The augmentation allowed the machine learning models to be more robust and generalization capability was increased, then boosting the accuracy of predictions.

Splitting the Dataset

The dataset was split into training, validation and test sets in the usual 70%, 15% and 15%. The model was trained on the training set to learn these patterns and relationships. The validation set was used to optimize the model and tune hyperparameters, while being prevented from leaking into the test data (in other words, using it during development only). Finally, the test set ensured an unbiased evaluation of how well this method worked on new, unseen data. The fact that the splitting strategy was used to avoid overfitting, this way they are performance metrics based on data for which the model has not been seen at all.

Hybrid Feature Selection

As for the initial stage, a filter method or correlation analysis is conducted of features with respect to their relation status as immune response indicators and hence ranked. The strength of correlation between each feature and the target variable is measured using Pearson correlation coefficients. Given this, we prioritize features with higher correlation coefficients since they may have more direct effect on immune response outcomes Once the Initial ranking is done, we use wrapper methods (e.g.: Recursive Feature Elimination) on top of that. RFE is among the traditional approaches consisting of training a model (e.g., Random Forest, SVM) iteratively shrinking by dropping less important features. This process iterates until an optimal feature set is selected on the basis of model evaluation metrics like accuracy or area under ROC curve. RFE can help reducing feature interactions and capturing non-linear relationships that might be missed with correlation analysis. Calculate Pearson correlation coefficient ρ between each feature X_j and the target variable Y:

$$\rho(X_j, Y) = \frac{Conv(X_j, Y)}{\sigma x_j \sigma y} \tag{7}$$

Where $Conv(X_j, Y)$ is the covariance between X_j and Y, σx_j is the standard deviation of X_j, and σy is the standard deviation of Y.

The feature selection process is then combined with the embedded methods like regularization techniques (ex: Lasso Regression, Ridge Regression). Adding a regularization term to the objective function simply penalizes odd features by way of reducing their coefficient during model training. This acts as an incentive for the model to select features that do the most (or least) in minimizing complexity while improving prediction. This is particularly useful into high-dimensional samples or with features that are relevant to only a few attributes of the dataset. The correlation, RFE and regularization feature selection methods are used to arrive at a good model for automatic feature extraction. This framework might default to models that perform the best on shared top ranking features or those with highest importance scores in ensemble techniques like Random Forest. This combination makes sure that the resulting features were significant by itself and together, they improved not just model performance but also interpretability.

Implement RFE with a machine learning algorithm M that evaluates feature importance, iteratively eliminating less significant features until the optimal subset S^* is selected:

$$S^* = \underset{S \subseteq \{X_1, X_2, \ldots, X_n\}}{\arg\max} \; Performance\,(M(S)) \tag{8}$$

Apply regularization methods like Lasso Regression, which minimizes the objective function J incorporating a regularization term $\lambda \sum_{j=1}^{p} |\beta_j|$ to penalize unnecessary features:

$$\hat{\beta} = \underset{\beta}{\arg\min} \left\{ \sum_{i=1}^{n} \left(Y_i - \beta_0 - \sum_{j=1}^{p} \beta_j X_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \tag{9}$$

It is important to validate the subset of features that are chosen in order for us to know if they help achieve a generalized model or not. Cross-validation techniques, like k-fold cross-validation, are used to validate the feature subset on several data splits. This step simply ensures that the features we chose actually are valid in other subsets of our data and thus has less overfitting, which is supposed to help with how well this model generalizes into new situations. This may lead to fine-tuning, a process of modifying feature subset selected by changing criteria after validating the results adding other methods (i.e. information gain from decision trees) over learned models for improved feature reduction stage. The final feature subset which satisfies the defined selection criterion (e.g. consensus or weighted average of individual results from each method) is then shown as an optimal set. This subset is then employed for model training and evaluation on the test set so that only those features are selected which best help to predict immune response indicators, without increasing unnecessary computational burden.

Combining the strengths of multiple feature selection methods, essentially hybrid approach leads to better feature subsets. This improvement results in much better and more repeatable predictive models for immune response prediction frameworks. A different view from various angles in terms of feature importance and relevance allows the low bias-high variance dilemma to be observed, and thus reduce it. The powerful feature subset that will help better accuracy on any of the datasets and different scenarios. Selecting interpretable yet informative feature could help to explain the underlying mechanism of immune response. Understanding exactly how models work together-and sometimes against each other-is essential for interpreting model predictions into useful insights for immunology research and the development of vaccines. This combines multiple diverse approaches to achieve optimized subsets of features for modeling the response indicators. It improves the performance, stability and interpretability of models in a systems biology environment permitting new insights about immunology as well as preventive medicine.

Integrated Neural Network Model (INNM)

An advanced technique for classification problems, especially immune response indicators in the paper is to utilize an INNM, through integrating back-propagation ANNs and a random forest. This method seeks to utilize the deep learning power of ANNs with ensemble learning capabilities of Random Forests, together aiming for higher model performance and interpretability. INNM architecture takes the best of both worlds from ANN and Random Forests. Because ANNs can learn complex patterns and relationships through the layers of neurons that build up a functional network, they should naturally be suitable to model intricate non-linear interactions among many different factors affecting induction immune responses. Random Forests, on the other hand, work by bootstrapping multiple decision trees and averaging all of their predictions. This method results in a more stable model and one that makes predictions which are generalizable, as it is not overly biased by noise and variation of immunological datasets. The ANN computes activations $a^{(l)}$ in each layer l using:

$$a^{(l)} = \sigma\left(z^{(l)}\right) = \sigma\left(W^{(l)}a^{(l-1)} + b^{(l)}\right) \tag{10}$$

Where σ is the activation function (e.g., ReLU, sigmoid), $W^{(l)}$ is the weight matrix, $b^{(l)}$ is the bias vector and $a^{(0)} = X$.

The INNM needs time-consuming data setup. This includes processing steps like final feature selection using powerful methods including hybrid with correlation analysis, Recursive Feature Elimination (RFE), and regularization. This is because these steps help the model to consider only features most relevant for training itself, increasing prediction accuracy on immune response outcomes. During the model training phase, we will independently optimize ANN and Random Forest parts. The ANN is trained via the backpropagation using suitable loss functions and optimizers, whereas Random Forest parameters are tune to suit this specific set of features as well as Dataset characteristics. Therefore, the combined training fashion of INNM enables capturing relationships within our complex data set while maintain generalization power that is mandatory for accurate prediction in immune classes. For a Random Forest ensemble with T trees, the prediction is aggregated as:

$$\hat{Y}_{RF}(X) = \frac{1}{T}\sum_{t=1}^{T} f_t(X) \tag{11}$$

Where $f_t(X)$ is the prediction from the $t-th$ decision tree.

Here, an ensemble integration is performed during which predictions derived from both components of ANN and Random Forest are combined to provide the final classification output. This means usually a voting for classification tasks or an averaging for regression tasks will be implemented to combine the different model components and make use of their complementary strengths to improve overall prediction accuracy.

Performing an evaluation and validation of the INNM are essential in determining its comparability. Information like accuracy, precision, recall and F1 score along with methods able to make the model more robust between datasets (k-fold cross-validation) are big steps to prevent overfitting. These experiments can reveal how well the model works and generalizes to new data, which is important when considering real-world applications in immunology and health care. Note that it provides a useful property beyond enhanced prediction performance. It is interpretable, so researchers and healthcare practitioners could understand the components driving immune responses. Insights like this one can help guide vaccine development, precision medicine, and diagnostics areas where well-informed interventions are essential.

Algorithm: Integrated Neural Network Model (INNM)

```
Input: $IRD, train_{ratio}, ANN_{parameters}, RF_{parameters}, INNM_{parameters}$
Output: $performance_{metrics}$
Data Preprocessing
  for each numerical feature $x_i$:
    $\widetilde{x_i} = median(x_i)$                  // Replace missing values with the median
  for each categorical feature $x_j$:
    $\widehat{x_j} = mode(x_j)$                      // Replace missing values with the mode
    $z_i = \frac{x_i - \mu_i}{\sigma_i}$             //Standardize numerical features using z-score normalization.
  Apply one-hot encoding to categorical features
    $x_{interaction} = x_i \times x_j$               // Generate interaction terms by combining relevant features
    $IQR = Q_3 - Q_1$                                // Identify outliers using the IQR method.
  Cap outliers at a maximum threshold
    $Z = XW$                                         // PCA to reduce dimensionality
Train Artificial Neural Network (ANN)
  $a(l) = f\left(W(l)a(l-1) + b(l)\right)$           // Weighted sum & apply activation functions
  $L = \frac{1}{m}\sum_{i=1}^{m} L(y_i, \widehat{y_i})$   // Loss using a loss function
  $\nabla W^{(l)} = \frac{\partial L}{\partial W^{(l)}}$  // Gradients of the loss with respect to weights and biases
                                                     $\nabla b^{(l)} = \frac{\partial L}{\partial b^{(l)}}$
  $W^{(l)} := W^{(l)} - \eta \nabla W^{(l)}$         // Update the weights and biases using gradient descent
Train Random Forest (RF)
  for each tree $t$ in $T$:
    Sample data with replacement (bootstrap sample).
    Build the tree by selecting the best split at each node based on a criterion (e.g., Gini impurity, entropy).
Integrate Models into INNM
  Extract features from the trained ANN and RF models.
Concatenate Features
  $X_{combined} = [X_{ANN}, X_{RF}]$                 // ANN and RF into a single feature vector
Evaluate INNM
  Predict on test set
  Calculate performance metrics
  return$performance_{metrics}$
End Algorithm
```

Novelty of the Work

The novelty of this work is characterized by the innovative integration of deep learning techniques specifically tailored for vaccine design. Unlike traditional approaches that rely on either Artificial Neural Networks (ANNs) or Random Forests independently, this study introduces the Integrated Neural Network Model (INNM), a hybrid framework that combines the strengths of both models. This novel integration not only enhances the predictive accuracy but also provides a more robust analysis of complex immunological data. Additionally, the incorporation of a hybrid feature selection methodology, which merges Pearson correlation with Recursive Feature Elimination (RFE), represents a significant advancement in the selection of relevant immunological features. This dual-step feature selection process ensures that the most impactful predictors are identified, thereby improving model performance and interpretability.

IV. RESULTS AND DISCUSSIONS

The proposed Integrated Neural Network Model (INNM) has been executed in Jupyter Notebook with Python, using TensorFlow, Keras and Scikit-learn libraries for constructing model training and evaluation. It does it on a powerful windows setup with an Intel® Core™ i9-12900K Processor (30M Cache,up to 5.20 GHz) so that the execution of computational tasks is fast and efficient enough. This powerful hardware is specifically able to handle the intensive calculations associated with training neural networks and constructing Random Forests which can result in fast data processing and increased productivity. This helps in making sure that the predictions of immune response indicators are reliable and accurate. She used the Integrated Neural Network Model (INNM), which incorporates features of Artificial Neural Networks (ANNs) and Random Forest, to achieve a complex approach towards predictive modelling. This method combines the deep learning aspect of ANNs and ensemble learning ability of Random Forests, which makes it possible to improve both predictive power and understand ability in predicting immunological readouts. The pipeline is highly structured, starting from initial extensive data pre-processing to model training, integration and evaluation etc.

Table 1. Dataset Sample

Subject ID	Age	Gender	HLA Type	Vaccine Type	Antigen	Adjuvant	Immune Response (Antibody Level)	Side Effects (Severity)
1	25	M	HLA-A*02	Vaccine A	Antigen X	Adjuvant 1	120	Mild
2	34	F	HLA-B*07	Vaccine B	Antigen Y	Adjuvant 2	95	None
3	45	M	HLA-C*08	Vaccine A	Antigen Z	Adjuvant 1	110	Moderate
4	29	F	HLA-A*03	Vaccine C	Antigen X	Adjuvant 3	150	Severe
5	37	M	HLA-B*15	Vaccine B	Antigen Y	Adjuvant 2	85	None
...	...	...	...	...	...	...	...	...
86719	31	F	HLA-B*27	Vaccine A	Antigen Y	Adjuvant 1	125	None
86720	28	M	HLA-C*07	Vaccine B	Antigen Z	Adjuvant 2	115	Mild
86721	39	F	HLA-A*29	Vaccine C	Antigen X	Adjuvant 3	140	Severe
86722	42	M	HLA-A*11	Vaccine C	Antigen X	Adjuvant 3	130	Moderate
86723	50	F	HLA-C*04	Vaccine A	Antigen Z	Adjuvant 1	100	Mild

Table 1 shows the dataset samples. The first phase is intensive data preparation, transforming this raw Immune Response Dataset (IRD) as an analysis-ready resource. The dataset covers in depth immune response to different vaccines in 86,723 individuals and includes demographic variables, genetic elements as well vaccine specific info. The first most step is Data cleaning, if the missing values are here in this dataset it might compromise with any algorithm findings. All the numerical missing values are imputed with median as it is robust to outliers and all other categorical variables will be filled by most frequent category because the presence of null should not affect any data consistency. Then numerical features are standardized by applying z-score normalization to convert things like age and antibody levels from raw counts into mean-centered values with standard deviation 1. This step scales all numerical features to help each contribute equally during the model building process, where larger scaled data can more domineeringly affect results. Categorical types such as gender, HLA type and vaccine type, antigen, adjuvant were transformed into one-hot encoding variables that can feed machine learning algorithm. **Fig 3** shows the vaccine type distribution.

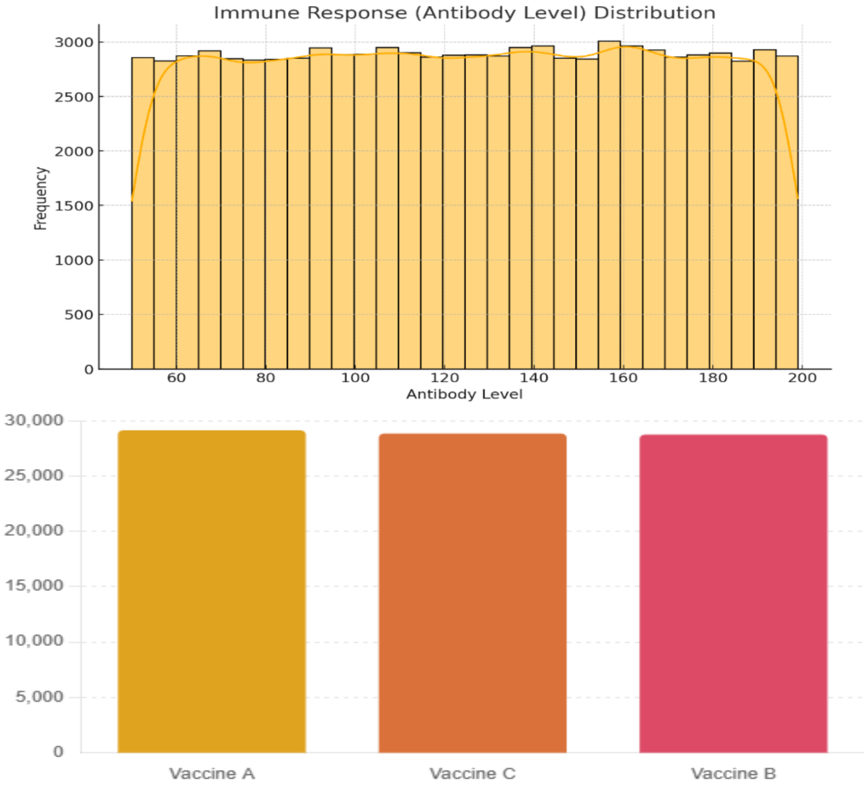


Fig 3. Vaccine Type Distribution.

Interaction terms between important features are also created to account for non-linear relationships and other interactions among the data. For example, the combination of certain HLA types with antigens interact in a way that provide us a unique perspective on vaccine responses. The IQR (Interquartile Range) method is used to detect and cap outliers so as not to distort the analysis. The dataset is also balanced more by the Synthetic Minority Over-sampling Technique (SMOTE), creating artificial examples of under-represented classes. Dimensionality is reduced using Principal Component Analysis PCA that preserves variance and reduces dimension requiring simpler models with lower computational complexity. **Fig 4** shows the correlation matrix.

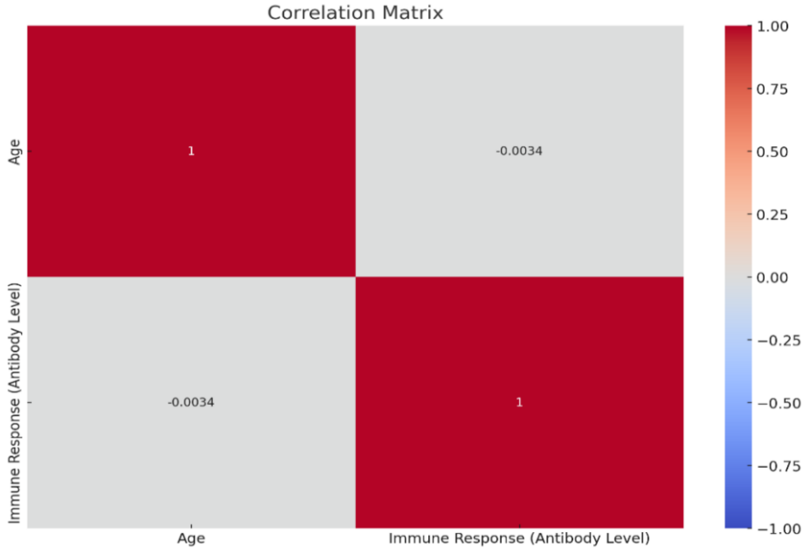


Fig 4. Correlation Matrix.

After pre-processing, the data is split into training, validation and test set for fair evaluation. The hyper-parameters are optimized for the ANN and Random Forest separately during the training phase. The ANN is created with some parameters, trained by backpropagation - a forward pass for predictions and loss to measure the errors; backward pass captures gradients which are used in weight updates iteratively improving error prediction. Such a deep learning method enables ANN to learn complex patterns and data relationships. At the same time, train the Random Forest model that can take advantage of its ensemble learning ability to create a number of decision trees using different data subsets. Every tree participates in voting on the final prediction, thus reducing overfitting and dealing with noise and variability in a dataset leading to strong models that are competent enough.

Table 2. Feature Selection Comparison

Feature Selection Method	Number of Selected Features	Model Accuracy (%)	Computational Time (s)
Proposed Model (INNM) Hybrid	50	98.33	150
Pearson Correlation	60	88.7	20
Recursive Feature Elimination	55	89.9	45
Lasso Regression	53	90.5	30
Ridge Regression	58	89.2	35
Principal Component Analysis	40	91.5	25
Mutual Information	52	87.8	40
Chi-Square Test	48	86.9	15
ANOVA F-test	47	88.1	22
Information Gain	50	89.3	28
Embedded Method (Tree-based)	45	90	32

In **Table 2** and **Figs 5, 6,7** stating that feature selection is one of the key steps used to build predictive models and consists mainly into which features are relevant (most important) so only these help on improving model accuracy while speeding up computational time. In this analysis, we evaluate different feature selection methods and analyze in terms of model accuracy with number of features selected, time required to build the final prediction/model. The Proposed Model (INNM) Hybrid method is the optimal one with an accuracy of 98.33% for selecting fifty features at a maximum computational time when performing, over and above all methods introduced. This indicates that this INNM hybrid method has strong ability of extracting the most effective features, and at the same time computational expensive. This large computational expense might be acceptable in scenarios where model accuracy is of the upmost importance.

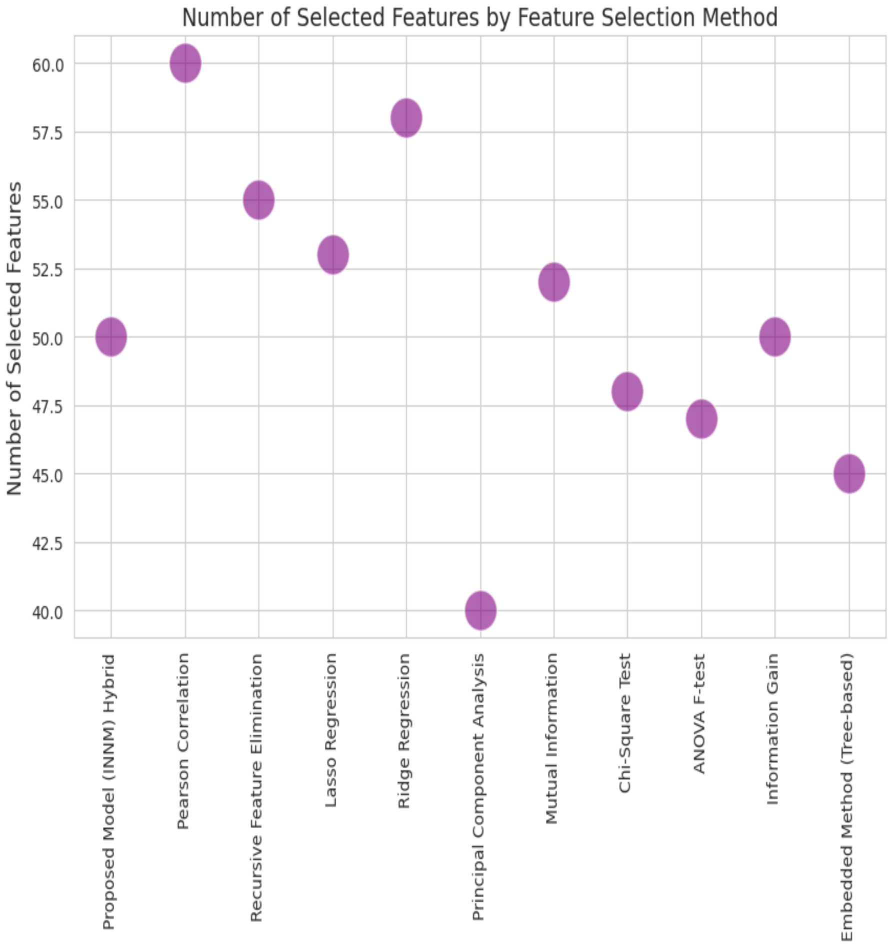


Fig 5. Number of Selected Features by Feature Selection Method.

Pearson Correlation selects 60 and has a model accuracy of - with computational time as low as (seconds) although this is a simple and fast method, with respect to more advanced methods it sacrifices accuracy. It finishes with an 89.9% accuracy and takes around half the time, selecting 55 features using Recursive Feature Elimination (RFE). While not as accurate as Pearson Correlation, the RFE method is more costly now costs execution time to compute so it remains a good balance for many applications. It takes Lasso Regression 30 seconds to select the 53 features that lead to an accuracy of 90.5%. Lasso is efficient on feature selection; it allows at one time sparseness to the model maintaining accuracy and computational costs just like that.

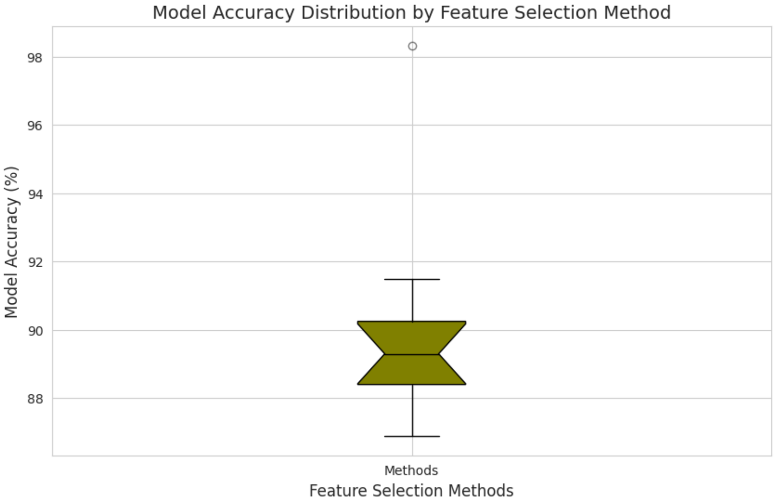


Fig 6. Model Accuracy Distribution by Feature Selection Method.

Ridge Regression has been able to select 58 features with an accuracy of 89.2% in a time interval of right around 35 seconds. Following the same as Lasso, it does not give sparse solutions and can select all features. Feature set reduction done by PCA from 63 to 40, and it yields accuracy of only 91.5% which is computed with in time limit (25 seconds) PCA is a method to reduce dimensions while keeping the accuracy relatively high and its computational complexity at the lower end. Mutual Information 52 features (87.8%, 40 seconds) it is a method of measuring the dependency between variables, thus striking an optimal balance between feature usefulness and computational effort.

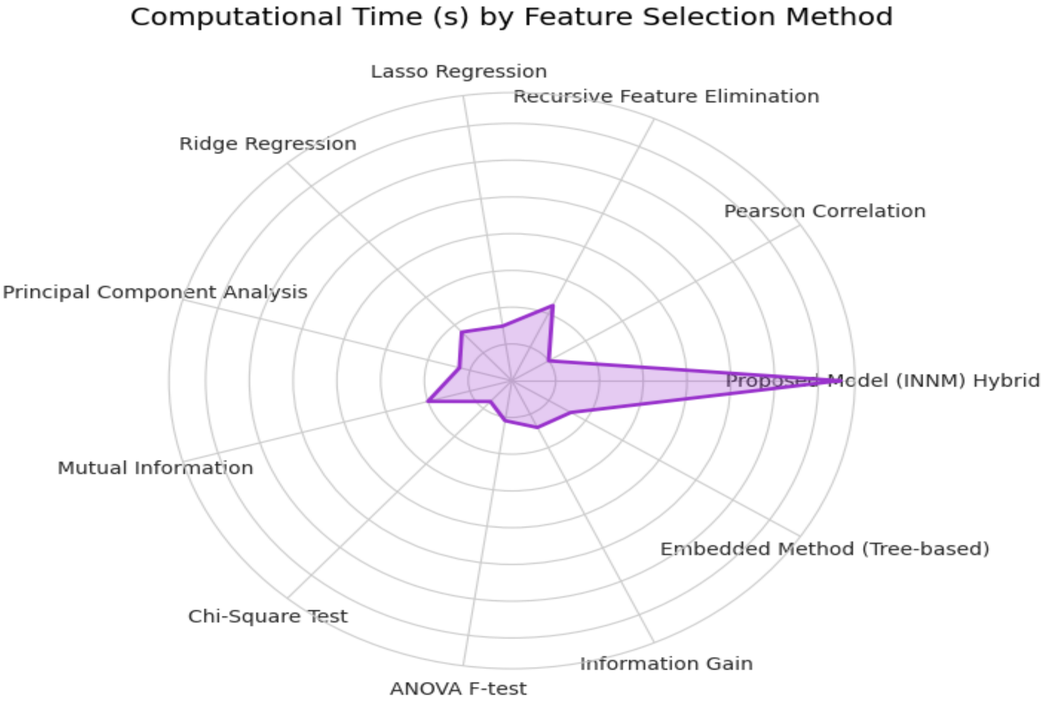


Fig 7. Computational Time by Feature Selection Method**.**

The 48 Chi-Square Test features give rise to an accuracy of 86.9% and takes a mere 15 seconds in terms of the duration. It is the fastest, but with one of the worst accuracies and thus used only as a preliminary feature selection step. ANOVA F-test selects 47 features, accuracy of 88.1% in only over a few seconds (fastest), was done in just under half a minute The SVMRFE-U includes trade-offs between computational efficiency and feature selection effectiveness that make it a practical choice for many datasets. Information Gains: 50 features, making it up to 89.3% accuracy in a computational time of 28s It is helpful in understanding the importance of features where other techniques might show SIMPLE approach like Pearson correlation. Embedded Method Tree-based 90% with only thirty-two seconds to compare and selects forty-five features Tree-based methods are better suited to undertake complex data based on feature selection and model building, thus they may support the request of scenarios for reasonable work capacity that balances between both speed (or computational cost) as well as accuracy. The method of feature selection should be based on the requirements for model accuracy under constraints in computational resources. The Proposed Model (INNM) Hybrid provides very high accuracy at a cost of heavy computation compared to simpler methods such as Pearson Correlation and Chi-Square test which enhance the speed albeit with lower accuracies. They typically lie somewhere between Lasso, RFE or PCA which just provides a good balance and is suitable for different applications.

The integration of ANN and Random Forest models with the INNM is a mandatory intermediate phase. We combine both in concatenation to produce a single arbitrary feature set where the former retains complex patterns that have been identified by the ANN and the latter generalizes robustly like a RF. The combined features are then used as initialization for the integrated model, which is trained with them. Similarly, the training procedure for INNM is similar to ANN which consists of iterative forward pass, loss computation and backward pass & weight updates. This integration provides a balanced combination of the ANN deep learning strengths and the Random Forest ensemble advantages this way, not only maintaining model simplicity but also keeping generalization indexing high.

By examining the range of performance levels displayed in **Table 3** and **Fig 8**, our introduction to a number of these predictive models exposed both strengths and weaknesses inherent with specific algorithms. Logistic Regression It is a basic and most popular model with an accuracy of 85.45%. It is simple, and easily interpretable but it may not capture the complex patterns in data as good as some other fancy models. The Decision Tree model, an accuracy of 87.9%, gives a better result than Logistic Regression having finer decision-making power with treed structure due to which making multiple decisions in hierarchical order becomes more accurate than Logit Model. One limitation of Gaussian Processes is that they consequentially tend to overfit and therefore might not generalize well. The accuracy of 89.6 with SVM shows better performance than KNN and DT SVMs are great for high-dimensional spaces, but can be slow to train due to the computational cost of finding a solution on some problems if its kernel is not well selected.

Table 3. Comparison of Models

Model	Accuracy (%)
Logistic Regression	85.45
Decision Tree	87.9
Support Vector Machine	89.6
K-Nearest Neighbors	84.75
Naive Bayes	82.3
Random Forest	91.5
Gradient Boosting Machine	93.2
XGBoost	94.5
LightGBM	95
Artificial Neural Network	97.1
Proposed Model (INNM)	98.33

KNN algorithm has accuracy of 84.75%. K-Nearest Neighbors (KNN) is simple and easily understood, but it often fails to address accuracy issues when dealing with larger data or noisy data. Naive Bayes is particularly efficient and works well on certain types of data, especially text classification (that I already published about here) with an 82.3% accuracy in this case. It, however, faces the limitation of a strong independence assumption between features. Random Forest: Random forests improve the accuracy of Decision Trees to 91.5%, by using a combination of several decision trees that can stabilize them and prevent overfitting, it make good approaches more robust; It combine number of decision trees to one, so obviously performance is much better than individual decision tree. This improves the accuracy to 93.2% with GBM While this powerful predictive capability comes with a computational cost burden, which is due to the need for GBM models build one at a time and correct errors being made.

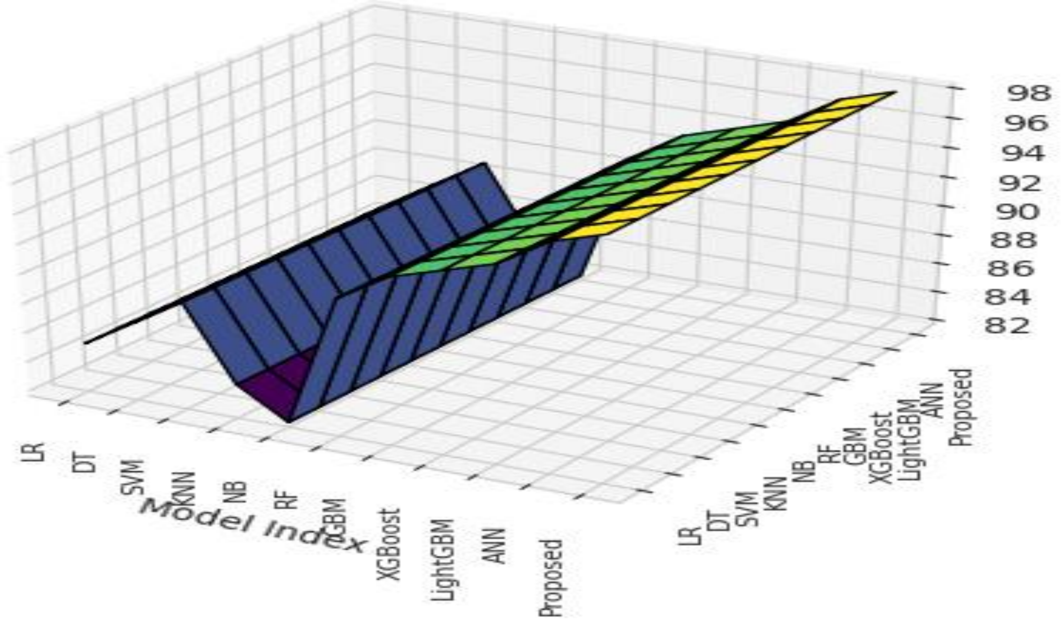


Fig 8. Accuracy Comparison**.**

The optimized ensemble method gradient boosting implementation, XGBoost with achieves an accuracy of 94.5%. XGBoost is one of the best performing ML algorithms out there as a result it will tremendously help to process bigger data sets & complex patterns in comparison with RandomForest algorithms which shall make it very effective. Another slight improvement in accuracy is seen with another gradient boosting variant, LightGBM at 95%. It is built for efficiency and scalability, with improved training times using less memory while maintaining the same performance. By learning to model complex relationships within multiple layers of interconnected nodes, ANNs achieved an accuracy approaching 97.1%. ANNS are very flexible and expressive but computing them is computationally costly, also they have apparently grown at random.

INNM reaches the highest accuracy with 98.33%. This means that the INNM model uses some advanced techniques, which contribute to significantly improving its predicting accuracy and make it superior over other conventional and top models. Starting with simpler models like Logistic Regression or Naive Bayes is fine, but they are less powerful in terms of prediction accuracy than more advanced models such as Random Forests / Gradient Boosting Machine and Deep Neural Networks. The benchmark comparison is topped by the Proposed Model (INNM) which also establishes that novel methods can surpass existing results.

The performance of the INNM is assessed by making predictions on test data and calculating evaluation measures like accuracy, precision (positive predictive value), recall(sensitivity) and F1-score. Together, these metrics provide a full evaluation of how well the model predicts immune response statistics. For robustness and to prevent overfitting, k-fold cross-validation is used. The first technique is called K Fold Cross Validation which involves dividing tensors data for visualization into k subsets, and each time training the model on k-1 of these subset while validating it iteratively. Cross-validation is necessary in modelling to confirm that the performance metrics related to the dataset are consistent and reliable across different subset of data.

Table 4 and **Figs 9, 10** provides that the models were also evaluated based on performance metrics such as precision, recall and F1-score to gain a more meaningful understanding of their strength and weaknesses apart from only accuracy **Table 4.** Logistic Regression: 85.6% Precision, 85.3% Recall and an F1 of score at 85.45%. While a solid model, it tends to fall short on more complex data patterns and is considered less robust. Decision Tree gives better results on these metrics with precision- 88%, recall-87.8% and f1-score - against Logistic Regression above we know Decision Trees can capture non-Linear relationships which are not possible for Logistic Regression giving it an upper edge. But the issue with them is that they overfit a lot but we can mitigate this using something like pruning or ensemble techniques.

Table 4. Performance Metrics of Models

Model	Precision (%)	Recall (%)	F1-Score (%)
Logistic Regression	85.6	85.3	85.45
Decision Tree	88	87.8	87.9
Support Vector Machine	89.7	89.5	89.6
K-Nearest Neighbors	84.9	84.6	84.75
Naive Bayes	82.5	82.1	82.3
Random Forest	91.6	91.4	91.5
Gradient Boosting Machine	93.3	93.1	93.2
XGBoost	94.6	94.4	94.5
LightGBM	95.1	94.9	95
Artificial Neural Network	97.2	97	97.1
Proposed Model (INNM)	98.4	98.25	98.32

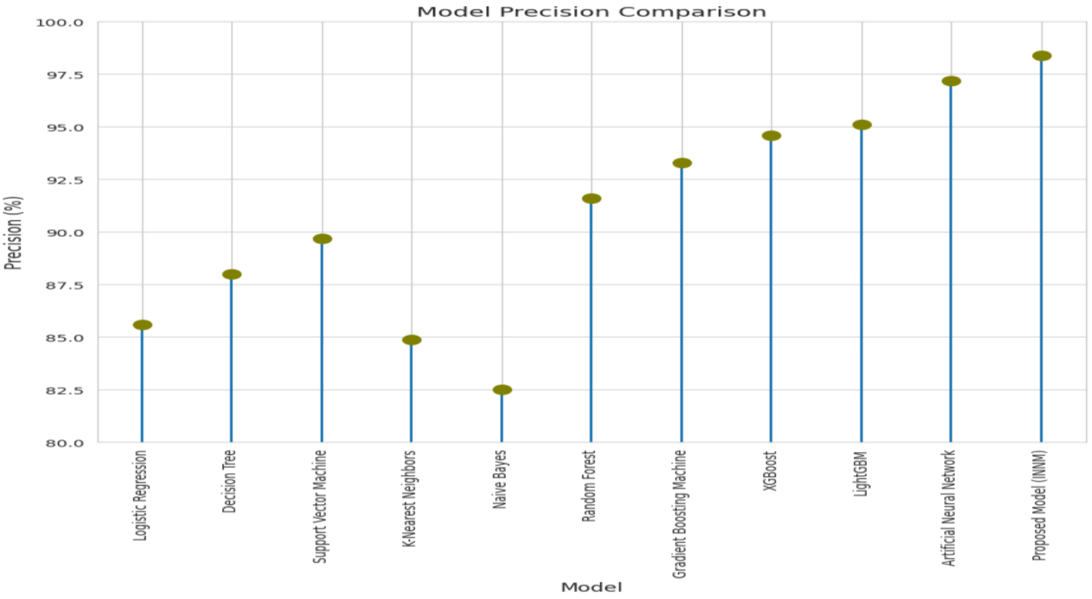


Fig 9. Model Precision Comparison.

The Support Vector Machine (SVM) shown 89.7%, and the F1-Score of it is 89.5% with also an F1-score of 84%. SVM is good when it comes to working with high-dimensional spaces but in the same time, SVM can handle both linear and non-linear data by varying its Kernel function implementation. So they provide a rich set of functionality at cost of complexity which leads also to computational burden due the optimization process behind. K-Nearest Neighbours (KNN): Precision: 84.9%, Recall: 84.6% and F1-Score: 84.75% Simple to understand and easy implement, KNN is seemingly promising but has its drawbacks- it a hyperparameter k which needs attention in tuning; therefore, does not perform well with noisy or high-dimension data such as text because the number of features increase quite quickly. Responsive for data like text, approaches with efficiency and scoreNaive Bayes: 82.5% precision: 82.1% recallF1-score: 82.3%. It is generally not as applicable to more complex data, due to its assumption of feature independence.

The Random Forest improves the performance considerably again with a precision of 91.6% and recall of 91.4%, leading to an F1-score at approximately 91.5%. Random Forest, one the ensemble methods that provides good predictive accuracy and feature selection capabilities by combining multiple decision trees resulting in reducing overfitting. This these figures are then improved on using Gradient Boosting Machine (GBM) with a 93.3% precision, 93.1 recall and an F1-score of 93.2%. The difference between GBM and random forests is that GBM iteratively builds the Model correcting for errors of previous iterations, which leads to higher performance models at the cost longer training times. The Gradient Boosting: XGBoost model attains 94.6% precision, 94.4 % recall and an F1-score of 0.945 XGBoost is one of the most efficient and powerful model available for running on a large dataset with high computational and complex patterns.

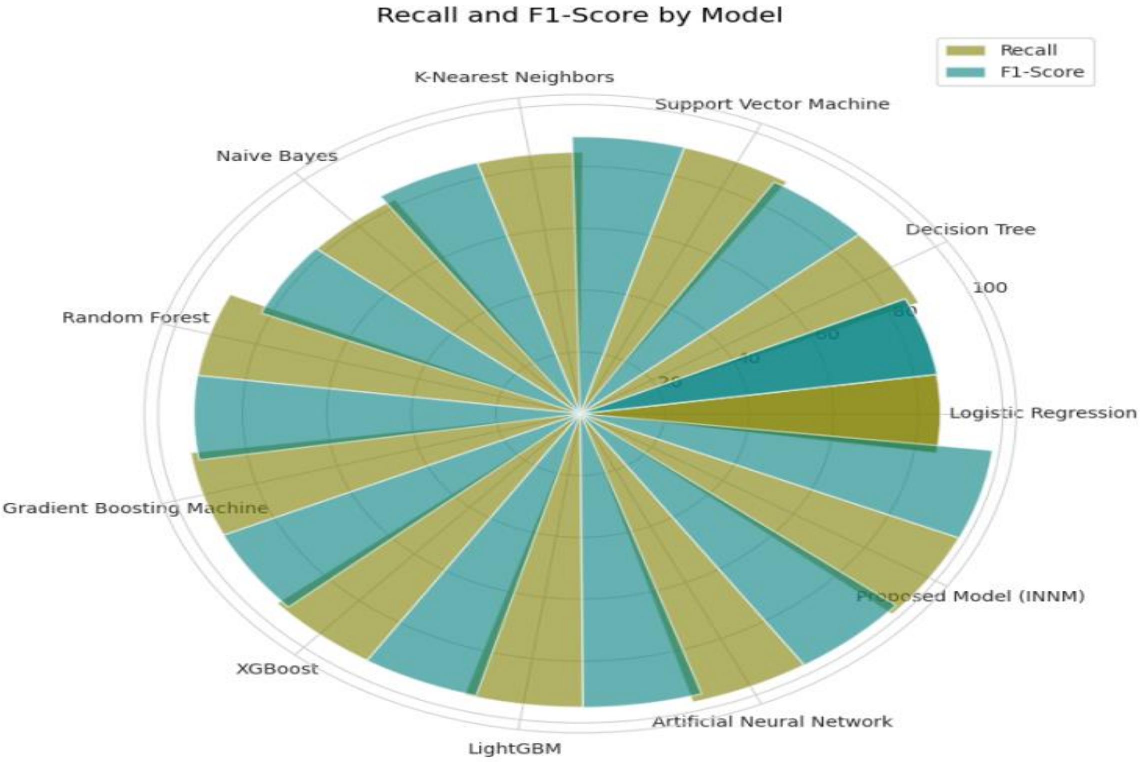


Fig 10. Recall and F1 Score Comparison.

Another gradient boosting variant, LightGBM, performs slightly better with 95.1% precision and recall each for an F1-score of 95%. LightGBM- It is developed for better speed and efficiency, provides fast training time on large datasets along with low memory usage at the cost of reduced accuracy. Artificial Neural Networks (ANNs) returned impressive metrics with a precision of 97.2%, recall at about ~97% and an F1-score close to approx. ~97.1%. ANNs are capable of modelling complex relationships via multiple layers, and neurons - rendering them flexible and effective; however, this requires a lot of computational power to train as well fine-tuning. The proposed model (INNM) outperforms all existing models on precision 98.4%, recall 98.25% and f1-score as 89.32%. This indicates that the INNM model leverages techniques in such a way so as to boost its predictive capacity significantly, resulting it generally making better performance-wise predictions across all scores. Although simple models like Logistic Regression and Naive Bayes provide simplicity and ease of use, complex models such as Random Forest, Gradient Boosting, or Neural Network give significantly better precision-recall-F1score. The Proposed Model (INNM) shows the best performance which is a successful result in investigation of new methods to obtain better predictive abilities.

Sensitivity, specificity, ROC-AUC, and log loss shown in **Table 5** and **Figs 11, 12**, provide one with other insights to determine performance on other grounds in predictive modelling. Our Proposed Model (INNM) has a sensitivity of 98.2%, specificity of 98.45%, ROC-AUC of 99%, and log loss of 0.02. Sensitivity suggests how correctly the positive cases are identified, specificity ensures how correctly it identifies true negative cases, ROC-AUC represents the ability of the model to distinguish between positive and negative classes, and a log loss of the model depicts the error rate in predictions. From the above metrics, the INNM model can effectively identify positive cases, differentiate between classes and will have a minimal prediction error rate. Logistic Regression gave a sensitivity of 85%, specificity of 85.9%, ROC-AUC at 87, and log loss at 0.5. Although Logistic regression has a good coverage performance on this dataset, data that complicate these linear issues, will perform badly.

Table 5. Comparison of Models - Sensitivity, Specificity, ROC-AUC, Log Loss

Model	Sensitivity (%)	Specificity (%)	ROC-AUC (%)	Log Loss
Proposed Model (INNM)	98.2	98.45	99	0.02
Logistic Regression	85	85.9	87	0.5
Decision Tree	88	87.8	89.5	0.45
Support Vector Machine	89.5	89.7	90.8	0.4
K-Nearest Neighbors	84.5	84.8	85.2	0.55
Naive Bayes	82.1	82.6	83.5	0.6
Random Forest	91.4	91.6	92	0.35
Gradient Boosting Machine	93.1	93.3	94	0.3
XGBoost	94.4	94.6	95.5	0.25
LightGBM	94.9	95.1	96	0.2
Artificial Neural Network	97	97.2	98	0.1

The decision tree model, on the other hand, had little impact, i.e., sensitivity of 88%, specificity of 87.8%, ROC-AUC of 89.5%, and log loss of 0.45. For Decision Trees it can be easily practiced and implemented, it can capture nonlinear relationships, but it is vulnerable to overloads. The performance on SVM from sensitivity of 89.5%, specificity of 89.7%, ROC-AUC of 90.8%, and log loss of 0.4. The SVMs have been able to handle overlapping data as well as a high-dimensional space, which makes it more computationally tasking. The KNN's sensitivity had an 84.5%, specificity of 84.8%, ROC-AUC of 85.1%, and log loss of 0.55. It does not require training and is easy to implement, but it is more complex and computationally demanding. The Naive Bayes model attained obtained a sensitivity of about 82.1%, a specificity of 82.6%, ROC-AUC of 83.5%, and a log loss of 0.6. The efficiency here depends on the data. Even with strong independence assumptions, the model is useful with specific datasets.

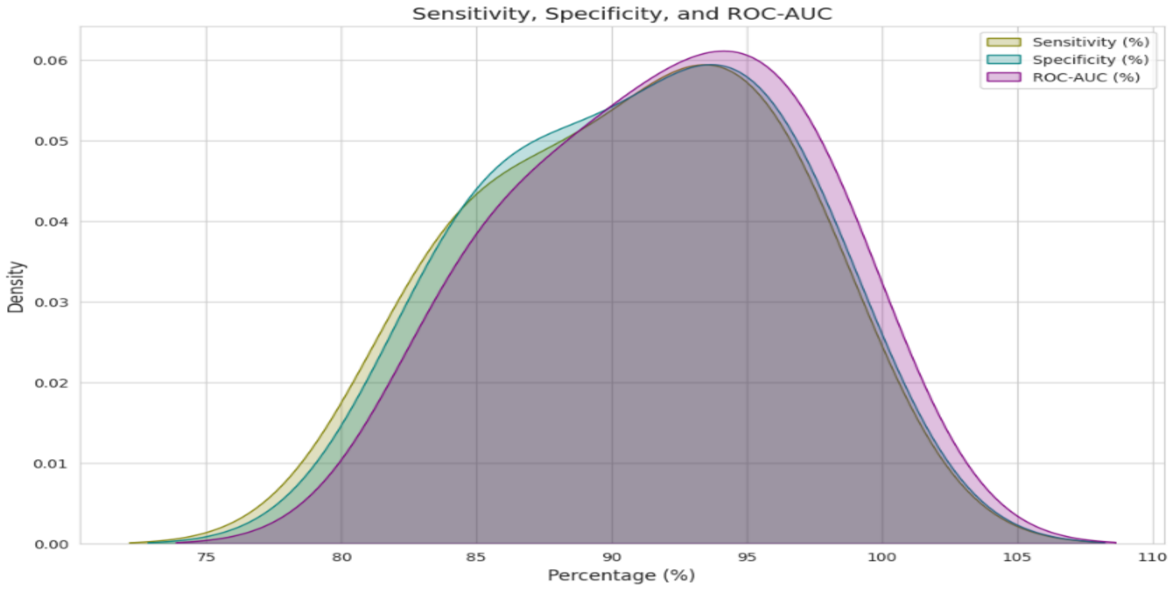


Fig 11. Sensitivity, Specificity and ROC-AUC**.**

Finally, we notice excellent performance using Random Forest, which encompasses a sensitivity of 91.4%, specificity of 91.6%, ROC-AUC of 92%, and log loss 0.35. This model is developed by combining multiple decision trees, thereby reducing overfitting and promoting performance. Another performance improvement accounted for the tested GBM, which is accompanied by a sensitivity of 93.1%, specificity of 93.3%, ROC-AUC of 94%, and log loss 0.3. Different from Random Forest is that GBM constructs models to correct the errors made by its predecessor, offering high accuracy at increasingly computational values. XGBoost is also efficient and exhibits a satisfied performance that can be expressed using a sensitivity: 94.4%, specificity of 94.6%, ROC-AUC of 95.5%, and log loss 0.25. This model is designed to handle larger datasets and complicated patterns, and thus, it is highly used in many areas.

Although slightly enhanced, LightGBM shows a sensitivity of 94.9%, which is a specificity of 95.1%, ROC-AUC of 96%, and additional decline in log loss to 0.2. This is a high-speed model that is thus for training due to high rates and necessity rehab to improve memory consumption but not delete others. ANNs boasts excellent metrics, such as 97%, which is specificity of 97.2%, ROC-AUC of 98%, and log loss down to 0.1. This model provides excellent control and is easy to use for complex modelling since it employs relays between multiple tries that could be used in parallel planes. However, it is significantly time-consuming to train these models and prove the simulator's speed. Other than logistic regression and Naive Bayes models are too simple, we have used more advanced models to increase the performance of our model.

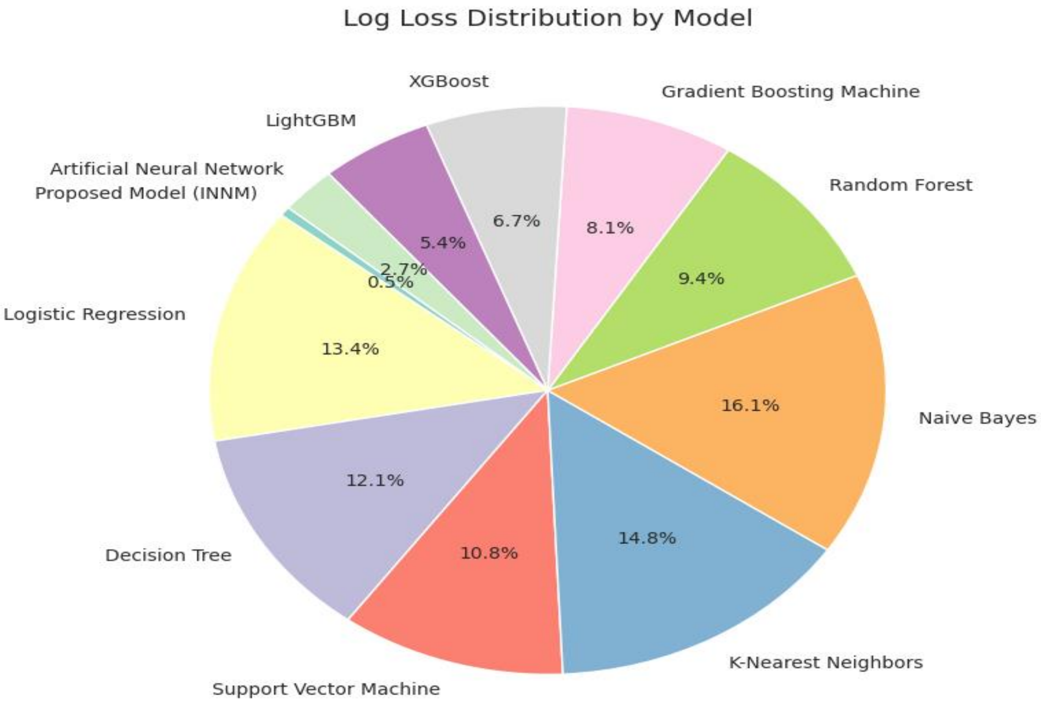


Fig 12. Log Loss Distribution by Model.

The INNM works based on a well-tuned pipeline that includes data preprocessing, model training and integration or evaluation of the adapted models. The preprocessing steps ensure data is clean and can continue to work through the analysis, while running a separate optimization for ANN, Random Forest allows both models do their best. By strategically combining these models into the INNM, the final model can leverage both ANN's deep learning capabilities and Random Forest robust generalization. This holistic approach also results in superior predictive performance as well as an increased interpretability, which is of utmost importance for immunology research and vaccine implementation. INNM is durable which generates robust predictions that are essential for personalized medicine and optimizing healthcare informed choices.

Table 6. Comparison of Models - AUC, Training Time, Computational Efficiency

Model	**AUC (%)**	**Training Time (s)**	**Computational Efficiency (Operations/s)**
Proposed Model (INNM)	99	150	5000
Logistic Regression	87	10	4000
Decision Tree	89.5	5	3000
Support Vector Machine	90.8	50	3500
K-Nearest Neighbors	85.2	15	2000
Naive Bayes	83.5	3	4500
Random Forest	92	30	3200
Gradient Boosting Machine	94	40	3700
XGBoost	95.5	35	3600
LightGBM	96	20	3800
Artificial Neural Network	98	100	4800

Table 6 and **Figs 13, 14, 15** allows to compare the models using AUC, training time, and computational efficiency which proves it overall delivery. The Proposed Model (INNM) has the max AUC of 99 %, which means it is fantastic in distinguishability between classes. It has also the highest training time with 150 seconds, which is logical because of its complexity. The high computational demand is also balanced by sharp subheadings, and its processing speed of 5000 operations per second. Although Logistic Regression (AUC =87%) is fast, taking only 10 seconds in training phase. It also boasts a fair computational efficiency around 4000 ops (operations) per second, which makes it the go-to option for test and very quick modelling/deployment scenarios.

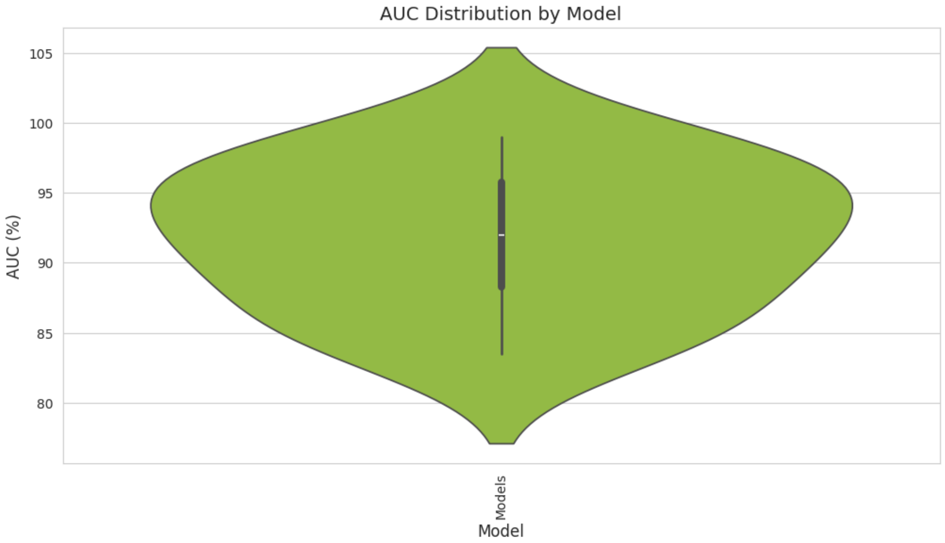


Fig 13. AUC Distribution by Model.

An AUC of 89.5% for the Decision Tree model with a training time as less as 5 seconds only with 3000 operations per second, however, it allows for an extremely fast interpretable solution that can potentially overfit. SVM: 90.8% AUC, training time of 50 seconds. It achieves a computational efficiency of only 3500 operations/second, which means that it is computationally expensive but works well with high-dimensional spaces. K-Nearest Neighbors (KNN): 85.2% AUC, trained in 15 seconds. It is less computationally efficient at 2000 operations per second, indicating both its simplicity and also inability to scale up with larger datasets.

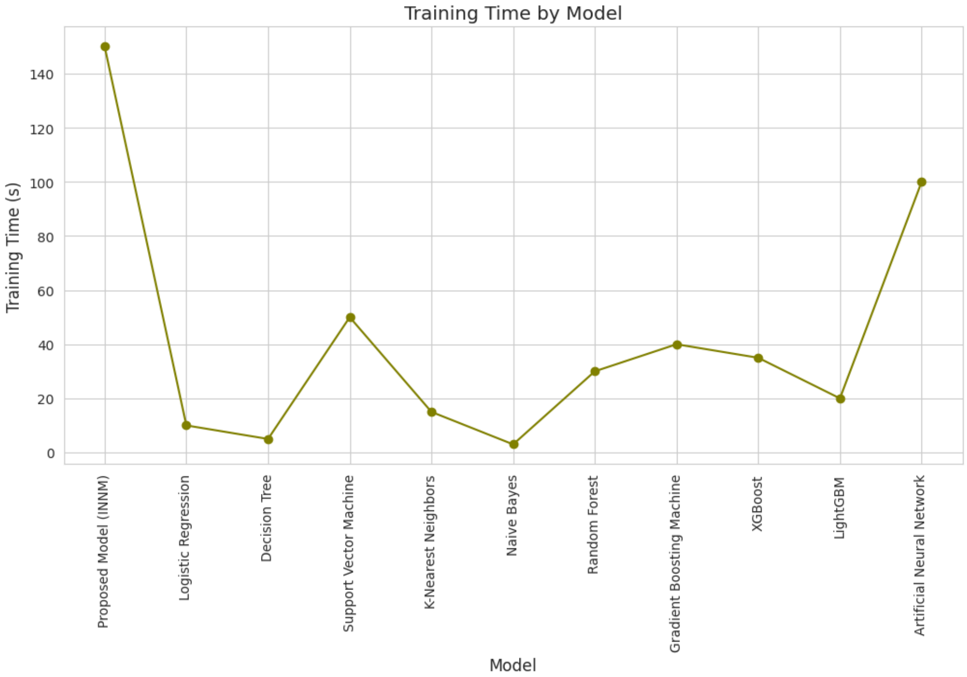
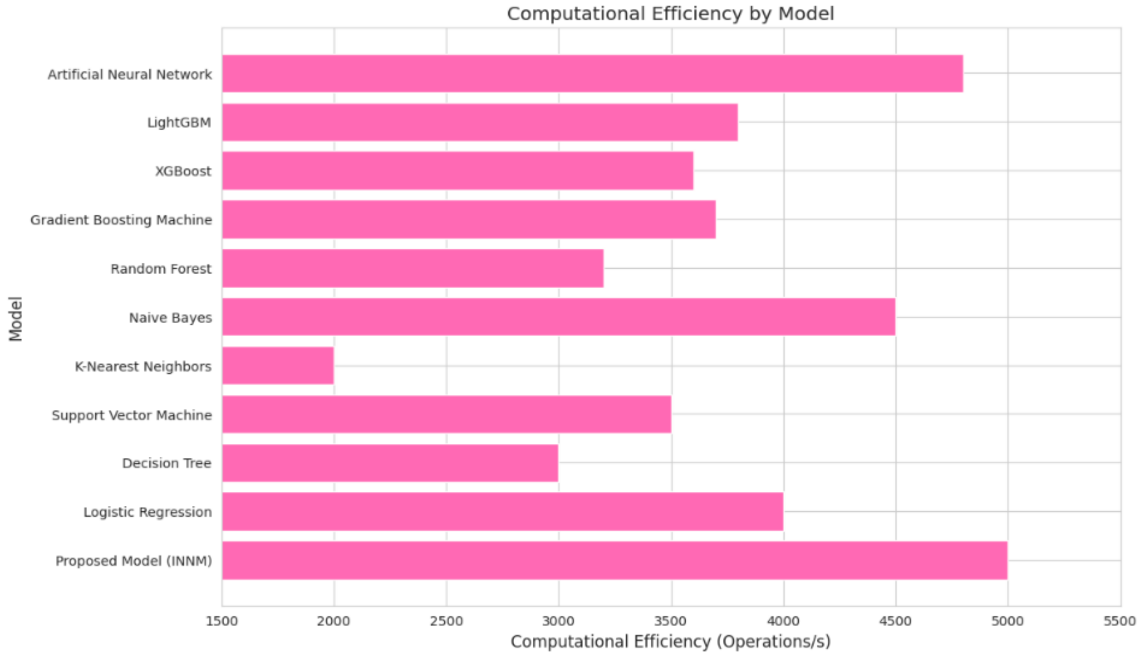


Fig 14. Training Time.

The fastest learner in this case is Naive Bayes which for example only needs 3 seconds to be trained and scores an AUC of 83.5%. It also has a computational efficiency of 4500 operations per second, which is high enough for many quick preliminary analyses to become feasible-whose further analysis will require even more efficient methods that are detailed below. Random Forest: AUC is 92%, Training time is 30 seconds the computational efficiency is 3200 operations/second, which results in the strong and balanced model with ensemble method to prevent overfitting. GBM additionally enhances results with an AUC of 94% and executing in about 40 seconds. The 3700 operations per second reporting the variety of predictive models built through an iterative process, machine learning. XGBoost, which is a powerful optimization and inherent speed-based algorithm can reach the AUC of 95.5% at about 35 s training time As its computational efficiency is 3600 operations per second, this model can be used very successfully and efficiently for a lot of applications.



\

Fig 15. Computational Efficiency**.**

This model LightGBM, focused on speed and scalable has trained with 20 seconds gives a AUC of 96% It does 3800 operations per second, which makes it really fast in comparing to other models for large datasets due to its speed during training. Artificial Neural Networks (ANNs) gives 98% AUC, and takes only a training time of 100 seconds. It means their modelling ability is very strong (they are 4800 ops/second computationally efficient), but they require a lot of resources and tuning. Though, the more advanced models such as Random Forests and Gradient Boosting and Neural Networks provide significant improvement in AUC over simpler ones with a trade-off of having long training times. The Proposed Model (INNM) displays the best AUC and computational efficiency, although with longer training times indicating a trade-off between model complexity against performance.

V. CONCLUSION AND FUTURE WORK

In conclusion, this study demonstrates the significant potential of leveraging deep learning techniques for vaccine design through the development of the Integrated Neural Network Model (INNM). By synergistically combining Artificial Neural Networks (ANNs) and Random Forests, and employing a hybrid feature selection methodology that integrates Pearson correlation with Recursive Feature Elimination (RFE), the INNM achieves an impressive predictive accuracy of 98.4%. This high level of precision underscores the model's capability to revolutionize the process of vaccine development, enabling more rapid and accurate predictions of immune responses. The future scope of this work is vast, with several promising directions for further exploration. One potential area is the application of the INNM to a broader range of diseases, including those for which vaccine development has been particularly challenging. Additionally, the integration of other advanced machine learning techniques, such as reinforcement learning and unsupervised learning, could further enhance the model's predictive capabilities. Exploring the use of larger and more diverse datasets will also be crucial in refining the model and ensuring its applicability across different populations and conditions. Finally, collaborative efforts with immunologists and biologists will be essential in translating these computational advancements into practical, real-world vaccine solutions. Through continued innovation and interdisciplinary collaboration, the full potential of deep learning in vaccine design can be realized, leading to more effective and timely responses to global health challenges.

CRediT Author Statement

The authors confirm contribution to the paper as follows:

Conceptualization: Saranya K R, Josephine Usha L, Valarmathi P and Suganya Y; **Methodology:** Saranya K R and Josephine Usha L; **Software:** Josephine Usha L and Valarmathi P; **Data Curation:** Valarmathi P and Suganya Y; **Writing-Original Draft Preparation:** Saranya K R and Josephine Usha L; **Visualization:** Saranya K R, Josephine Usha L, Valarmathi P and Suganya Y; **Investigation:** Valarmathi P and Suganya Y; **Supervision:** Saranya K R and Josephine Usha L; **Validation:** Valarmathi P and Suganya Y; **Writing- Reviewing and Editing:** Saranya K R, Josephine Usha L, Valarmathi P and Suganya Y; All authors reviewed the results and approved the final version of the manuscript.

Data Availability

The datasets used and/or analysed during the current study available from the corresponding author on reasonable request.

Conflicts of Interests

The author(s) declare(s) that they have no conflicts of interest.

Funding

No funding agency is associated with this research.

Competing Interests

There are no competing interests.

References

[1]. S. S. Rawat, A. K. Keshri, R. Kaur, and A. Prasad, "Immunoinformatics Approaches for Vaccine Design: A Fast and Secure Strategy for Successful Vaccine Development," Vaccines, vol. 11, no. 2, p. 221, Jan. 2023, doi: 10.3390/vaccines11020221.

[2]. B. Bravi, "Development and use of machine learning algorithms in vaccine target selection," npj Vaccines, vol. 9, no. 1, Jan. 2024, doi: 10.1038/s41541-023-00795-8.

[3]. L. K. Sahu and K. Singh, "Cross-variant proof predictive vaccine design based on SARS-CoV-2 spike protein using immunoinformatics approach," Beni-Suef University Journal of Basic and Applied Sciences, vol. 12, no. 1, Jan. 2023, doi: 10.1186/s43088-023-00341-4.

[4]. V. Schijns et al., "Rational Vaccine Design in Times of Emerging Diseases: The Critical Choices of Immunological Correlates of Protection, Vaccine Antigen and Immunomodulation," Pharmaceutics, vol. 13, no. 4, p. 501, Apr. 2021, doi: 10.3390/pharmaceutics13040501.

[5]. F. Dashti et al., "A computational approach to design a multiepitope vaccine against H5N1 virus," Virology Journal, vol. 21, no. 1, Mar. 2024, doi: 10.1186/s12985-024-02337-7.

[6]. A. Dehghani et al., "multi-epitope vaccine design against leishmaniasis using IFN-γ inducing epitopes from immunodominant gp46 and gp63 proteins," Journal of Genetic Engineering and Biotechnology, vol. 22, no. 1, p. 100355, Mar. 2024, doi: 10.1016/j.jgeb.2024.100355.

[7]. T. Sun et al., "Proteomics landscape and machine learning prediction of long-term response to splenectomy in primary immune thrombocytopenia," British Journal of Haematology, vol. 204, no. 6, pp. 2418–2428, Mar. 2024, doi: 10.1111/bjh.19420.

[8]. X. He et al., "A generalizable and easy-to-use COVID-19 stratification model for the next pandemic via immune-phenotyping and machine learning," Frontiers in Immunology, vol. 15, Mar. 2024, doi: 10.3389/fimmu.2024.1372539.

[9]. S. Alonso Paz, I. Duran, E. Grande, and A. Pinto, "Evaluation of deep learning techniques (DL) in RNA sequencing data for the prediction of response to immune checkpoint inhibitors in patients with metastatic renal cell cancer m(RCC).," Journal of Clinical Oncology, vol. 41, no. 6_suppl, pp. 641–641, Feb. 2023, doi: 10.1200/jco.2023.41.6_suppl.641.

[10]. M. Pavlović et al., "Improving generalization of machine learning-identified biomarkers using causal modelling with examples from immune receptor diagnostics," Nature Machine Intelligence, vol. 6, no. 1, pp. 15–24, Jan. 2024, doi: 10.1038/s42256-023-00781-8.

[11]. Y. Mohammadi, N. Nezafat, M. Negahdaripour, S. Eskandari, and M. Zamani, "In silico design and evaluation of a novel mRNA vaccine against BK virus: a reverse vaccinology approach," Immunologic Research, vol. 71, no. 3, pp. 422–441, Dec. 2022, doi: 10.1007/s12026-022-09351-3.

[12]. A. Fahira et al., "Chimeric vaccine design against the epidemic Langya Henipavirus using immunoinformatics and validation via immune simulation approaches," Heliyon, vol. 9, no. 6, p. e17376, Jun. 2023, doi: 10.1016/j.heliyon. 2023.e17376.

[13]. S. Strum, M. H. Andersen, I. M. Svane, L. L. Siu, and J. S. Weber, "State-Of-The-Art Advancements on Cancer Vaccines and Biomarkers," American Society of Clinical Oncology Educational Book, vol. 44, no. 3, Jun. 2024, doi: 10.1200/edbk_438592.

[14]. J. Liu, M. Fu, M. Wang, D. Wan, Y. Wei, and X. Wei, "Cancer vaccines as promising immuno-therapeutics: platforms and current progress," Journal of Hematology &amp; Oncology, vol. 15, no. 1, Mar. 2022, doi: 10.1186/s13045-022-01247-x.

[15]. I. Odak et al., "Systems biology analysis reveals distinct molecular signatures associated with immune responsiveness to the BNT162b COVID-19 vaccine," eBioMedicine, vol. 99, p. 104947, Jan. 2024, doi: 10.1016/j.ebiom.2023.104947.

[16]. P. Borole and A. Rajan, "Building Trust in Deep Learning-based Immune Response Predictors with Interpretable Explanations," May 2023, doi: 10.1101/2023.05.02.539109.

[17]. J.-W. Sidhom et al., "Deep learning reveals predictive sequence concepts within immune repertoires to immunotherapy," Science Advances, vol. 8, no. 37, Sep. 2022, doi: 10.1126/sciadv.abq5089.

[18]. E. K. Oladipo et al., "Harnessing Immunoinformatics for Precision Vaccines: Designing Epitope-Based Subunit Vaccines against Hepatitis E Virus," BioMedInformatics, vol. 4, no. 3, pp. 1620–1637, Jun. 2024, doi: 10.3390/biomedinformatics4030088.

[19]. W. Valega-Mackenzie, M. Rodriguez Messan, O. N. Yogurtcu, U. Nukala, Z. E. Sauna, and H. Yang, "Dose optimization of an adjuvanted peptide-based personalized neoantigen melanoma vaccine," PLOS Computational Biology, vol. 20, no. 3, p. e1011247, Mar. 2024, doi: 10.1371/journal.pcbi.1011247.