Immunology Letters 184 (2017) 51–60

Contents lists available at ScienceDirect

Immunology Letters

journal homepage: www.elsevier.com/locate/immlet

# Multilevel ensemble model for prediction of IgA and IgG antibodies

Divya Khanna*, Prashant Singh Rana

*Computer Science and Engineering Department, Thapar University, Patiala, Punjab 147004, India*

ARTICLE INFO

*Article history:*
Received 22 November 2016
Received in revised form 30 January 2017
Accepted 30 January 2017
Available online 16 February 2017

*Keywords:*
B-cell epitope
Antibody
Machine learning models
Multilevel ensemble model
Regularized trees

ABSTRACT

Identification of antigen for inducing specific class of antibody is prime objective in peptide based vaccine designs, immunodiagnosis, and antibody productions. It's urge to introduce a reliable system with high accuracy and efficiency for prediction. In the present study, a novel multilevel ensemble model is developed for prediction of antibodies IgG and IgA. Epitope length is important in training the model and it is efficient to use variable length of epitopes. In this ensemble approach, seven different machine learning models are combined to predict variable length of epitopes (4 to 50). The proposed model of IgG specific epitopes achieves 94.43% of accuracy and IgA specific epitopes achieves 97.56% of accuracy with repeated 10-fold cross validation. The proposed model is compared with the existing system i.e. IgPred model and outcome of proposed model is improved.

## 1. Introduction

In immune response, immunoglobulins plays an important role i.e. they recognize the specific antigen and blind with it and hence protect the body. Antigen can be toxin, bacteria, and virus. When immune system is not work properly or it doesn't generate required antibodies is known as immunodeficiency. It can be through side effect of medicines, disease, infection or lack of proper nutrition. The analysis of particular immunoglobulins' presence in the blood can be useful for the identification of infections or different illnesses or to do the diagnoses. The types of antibodies are IgM, IgE, IgD, IgA and IgG that are generated by our immune system to fight against invaders. The amount of Immunoglobulin G (IgG) is huge, followed by Immunoglobulin M (IgM) and Immunoglobulin A (IgA) [1]. The amount of Immunoglobulin D (IgD) is less than IgA, IgG, IgM but higher than Immunoglobulin E (IgE) [2]. IgM and IgG preserve from infections in internal body tissues, blood and organs. IgA [3] is available in blood, the greater part of the IgA in the body is in the secretions of the mucosal surfaces which encompass respiratory, saliva, tears and gastrointestinal secretions. The IgA antibodies in the secretions play a major role in protecting from infections in these areas. IgG and IgM are also found in secretions but not equal to the amount of IgA. The human mucosal surfaces is protected by IgA that shield is very important for human body and deficiency of IgA may lead to cancer [4]. IgG the most abundant type of antibody, is found in all body fluids and protects against bacterial and viral infections [5]. In the target antigens, identification of B-cell epitopes is important to design epitope-driven vaccine, immunodiagnostic tests and production of antibody. B-cell epitopes can be classified into two classes: linear epitopes and discontinuous epitopes. Linear epitopes are peptides that have amino acid residues adjacent in the polypeptide chain. Discontinuous epitopes are created from amino acid residues located in different parts of the polypeptide chain. 90% of epitopes are discontinuous epitopes and rest are sequential epitopes. In experimental designs, immunodiagnostic tests and vaccines production [6], the disclosure of continuous epitopes still play a key role because most of the B-cell epitopes are discontinuous.

In the late years, numerous computational strategies have been produced to foresee B-cell epitopes. The propensity scales such as hydrophilicity [7], surface accessibility [8] and antigenicity [9] have been proposed to predict linear B-cell epitopes. The epitopes combination of physicochemical properties have been used to develop the methods such as PREDITOP [10], PEOPLE [11], BEPITOPE [12] and BcePred [13] for prediction of linear B-cell. The forecast exactness is expanded by utilizing physicochemical property as contrast with the techniques that utilizes the single property. ABCPred [14], Chen et al. [15], BCPred [16], SVMTrip [17], Huang [18], LIAN Yao [19], APCpred [20] utilize machine learning algorithms to predict linear B-cell epitopes. All these models are used to predict B-cell epitopes, not for identifying specific class of antibodies in antigen. Numerous methods have been developed for predicting antigenic

* Corresponding author.
*E-mail address:* divyakhanna1991@gmail.com (D. Khanna).

http://dx.doi.org/10.1016/j.imlet.2017.01.017
0165-2478/© 2017 European Federation of Immunological Societies. Published by Elsevier B.V. All rights reserved.

**Table 1**
Illustration of the features.

| Feature | Information |
|---|---|
| CL | Class i.e. IgG, IgA or Non-IgG, Non-IgA |
| $F_a$ | Aliphatic index |
| $F_b$ | Potential protein interaction index |
| $F_c$ | Hydrophobic moment |
| $F_d$ | Instability index |
| $F_e$ | Probability of detection of peptides |
| $F_f$ | Number of possible neighbours |
| $F_g$ | Tiny |
| $F_h$ | Small |
| $F_i$ | Aliphatic |
| $F_j$ | Aromatic |
| $F_k$ | Nonpolar |
| $F_l$ | Polar |
| $F_m$ | Charged |
| $F_n$ | Basic |
| $F_o$ | Acidic |
| $F_p$ | Percentage of tiny |
| $F_q$ | Percentage of small |
| $F_r$ | Percentage of aliphatic |
| $F_s$ | Percentage of aromatic |
| $F_t$ | Percentage of nonpolar |
| $F_u$ | Percentage of polar |
| $F_v$ | Percentage of charged |
| $F_w$ | Percentage of basic |
| $F_x$ | Percentage of acidic |
| $F_y$ | Charge of protein sequence |
| $F_z$ | Hydrophobicity |
| $F_{aa}$ | Kidera factor |
| $F_{ab}$ | Molecular Weight |
| $F_{ac}$ | Isoelectric point |
| SL | Sequence Length |

regions or B-cell epitopes that can induce B-cell response. The concentration is on the prediction of antibody like IgA, IgE and IgG. Algpred [21] predicts the allergenic proteins based on four different approaches and scores accuracy of 85.02%. First approach, support vector machine (SVM) is used to predict allergens based on amino acid and dipeptide composition of proteins. Second approach, motif based technique is used to predict allergens by using software MEME/MAST [22]. Third approach, segment similarity technique is used. If segment is similar to allergen representative proteins (ARPs) [23], the protein is assigned allergen. Fourth approach, If protein has segment similar to known IgE epitopes that protein is considered as allergen. Another author has used pseudo-amino acid composition (PseAAC) and SVM [24] to predict allergenic proteins (IgE). The efficiency of model is evaluated by using fivefold cross-validation and results are better than the previous work by scoring accuracy of 91.20%. Another method i.e. IgPred [25] predicts antigens of specific type of antibodies that utilizes the amino acid sequence information for prediction. The model is used to identify the specific class of antibodies in the antigen by using features like binary profiles, dipeptide composition and amino acid composition. From these features, dipeptide composition-based SVM achieved accuracy of 70.72%, 82.7% and 72.07% for IgG, IgE and IgA specific epitopes respectively. Models are evaluated using five-fold cross validation. They have provided the dataset of IgE, IgA and IgG antibodies and those are used in the present study.

Prediction of antibody specific class is important for testing immune system, finding out the allergy, infection and any other illness. Influenced from the outcome of machine learning approaches and the urge to find an accurate method which predict B-cell epitopes that can induce a specific class of antibody (like IgA, IgG) a novel multilevel ensemble model is proposed. The physicochemical properties of amino acid are used to train machine learning models that will classify the IgG and IgA inducing epitopes as explained in Section 3.

The paper is organized as follows: A hasty overview of the evaluated features, dataset, feature selection, machine learning models and benchmark dataset are presented in Section 2. The methodology and proposed model are explained in Section 3. Model evaluation is introduced in Section 4. Section 5 narrates experiments, result analysis, comparison and discussion. Finally, conclusion and future work is presented in Section 6.

## 2. Materials and methods

### 2.1. Dataset and its features

The balanced dataset consists of IgA and IgG inducing epitopes taken from http://crdd.osdd.net/raghava/IgPred/. Total number of IgG epitopes sequences are 16,067, in which IgG are 7,575 and non-IgG are 7,673; IgA epitopes are 403 and non-IgA are 416. All the sequences have variable length from 4 to 50. Table 1 shows the physicochemical properties used in the study. The glimpse of datasets are presented in Tables 2 and 3 .

### 2.2. Feature measurement

Table 4 describes the physicochemical properties of amino acid that are used in the present study. To extract the properties, an open source software R i.e. licensed under GNU GPL is used.

**Table 2**
XXX.

| SL | $F_a$ | $F_b$ | $F_c$ | $F_d$ | ––– | $F_z$ | $F_{aa}$ | $F_{ab}$ | $F_{ac}$ | $CL_{IgG}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 7 | 111.43 | 2.61 | 0.41 | 71.34 | ––– | -0.03 | -0.07 | 834.89 | 6.50 | 1 |
| 11 | 26.36 | 3.29 | 0.25 | 37.40 | ––– | -0.15 | -0.25 | 1246.21 | 3.23 | 1 |
| 15 | 90.67 | 0.09 | 0.13 | 54.03 | ––– | 0.31 | -0.67 | 1498.66 | 7.54 | 1 |
| 12 | 98.33 | -0.05 | 0.34 | -4.98 | ––– | 0.30 | -1.00 | 1088.23 | 6.41 | 0 |
| 6 | 131.67 | 0.46 | 0.45 | 8.33 | ––– | 0.22 | -0.93 | 571.68 | 10.55 | 0 |
| 10 | 98.00 | -1.20 | 0.16 | 19.77 | ––– | 0.39 | -1.00 | 886.08 | 8.55 | 0 |

**Table 3**
Sample dataset of IgA epitopes.

| SL | $F_a$ | $F_b$ | $F_c$ | $F_d$ | ––– | $F_z$ | $F_{aa}$ | $F_{ab}$ | $F_{ac}$ | $CL_{IgA}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 10 | 20.00 | 3.73 | 0.22 | 47.87 | ––– | -0.60 | -0.03 | 1176.35 | 9.44 | 1 |
| 5 | 118.00 | -0.82 | 0.35 | 8.00 | ––– | 0.44 | -0.59 | 533.64 | 3.85 | 1 |
| 20 | 29.50 | 2.61 | 0.40 | 53.96 | ––– | -0.33 | -0.08 | 2267.58 | 8.52 | 1 |
| 40 | 66.00 | 1.30 | 0.33 | 28.29 | ––– | 0.07 | -0.47 | 4212.79 | 8.24 | 0 |
| 12 | 137.50 | 0.32 | 0.31 | 11.07 | ––– | 0.19 | -0.39 | 1275.56 | 11.65 | 0 |
| 15 | 65.33 | 1.61 | 0.29 | 32.57 | ––– | -0.15 | -0.34 | 1630.90 | 8.23 | 0 |

**Table 4**
Physicochemical properties of amino acid.

| Sr. No. | Property | Description | R Package | Function |
|---|---|---|---|---|
| **1** | Aliphatic index | The relative volume occupied by aliphatic side chains (alanine, valine, isoleucine, and leucine) is known as aliphatic index of a protein. | Peptides [26] | aindex |
| **2** | Potential protein interaction index | Based upon amino acid sequence of a protein, the potential protein interaction index is computed i.e. introduced by Boman [27]. | Peptides | boman |
| **3** | Hydrophobic moment | It is computed for an amino acid sequence of N residues and their corresponding hydrophobicities | Peptides | hmoment |
| **4** | Instability index | The stability of protein in a test tube is estimated by instability index. | Peptides | instaindex |
| **5** | Probability of detection of peptides | Probability of peptides' occurrence. | Peptider [28] | ppeptide |
| **6** | Number of possible neighbours | It describes neighbors of degree one for a group of peptide sequences. | Peptider | getNofNeighbors |
| **7** | Tiny | Number of amino acid in the sequence which comes under tiny class. | Peptides | aacomp |
| **8** | Small | Number of amino acid in the sequence which comes under small class. | Peptides | aacomp |
| **9** | Aliphatic | Number of amino acid in the sequence which comes under aliphatic class. | Peptides | aacomp |
| **10** | Aromatic | Number of amino acid in the sequence which comes under aromatic class. | Peptides | aacomp |
| **11** | Nonpolar | Number of amino acid in the sequence which comes under nonpolar class. | Peptides | aacomp |
| **12** | Polar | Number of amino acid in the sequence which comes under polar class. | Peptides | aacomp |
| **13** | Charged | Number of amino acid in the sequence which comes under charged class. | Peptides | aacomp |
| **14** | Basic | Number of amino acid in the sequence which comes under basic class. | Peptides | aacomp |
| **15** | Acidic | Number of amino acid in the sequence which comes under acidic class. | Peptides | aacomp |
| **16** | Percentage of tiny | Percentage of tiny amino acid in the given sequence. | Peptides | aacomp |
| **17** | Percentage of small | Percentage of small amino acid in the given sequence. | Peptides | aacomp |
| **18** | Percentage of aliphatic | Percentage of aliphatic amino acid in the given sequence. | Peptides | aacomp |
| **19** | Percentage of aromatic | Percentage of aromatic amino acid in the given sequence. | Peptides | aacomp |
| **20** | Percentage of nonpolar | Percentage of nonpolar amino acid in the given sequence. | Peptides | aacomp |
| **21** | Percentage of polar | Percentage of polar amino acid in the given sequence. | Peptides | aacomp |
| **22** | Percentage of charged | Percentage of charged amino acid in the given sequence. | Peptides | aacomp |
| **23** | Percentage of basic | Percentage of basic amino acid in the given sequence. | Peptides | aacomp |
| **24** | Percentage of acidic | Percentage of acidic amino acid in the given sequence. | Peptides | aacomp |
| **25** | Charge of protein sequence | The net charge can be computed at defined pH by using pKa scales. | Peptides | charge |
| **26** | Hydrophobicity | It computes the hydrophobicity index of an amino acids sequence. | Peptides | hydrophobicity |
| **27** | Kidera factor | The Kidera Factors is derived by applying multivariate analysis to 188 physical features of the 20 amino acids and using dimension reduction techniques. | Peptides | kidera |
| **28** | Molecular Weight | This function calculates the molecular weight of a protein sequence. | Peptides | mw |
| **29** | Isoelectric point | The isoelectric point (pI), is the pH at which a particular molecule or surface doesn't carry electrical charge. | Peptides | pI |

### *2.3. Feature importance using regularized trees*

While building model, feature selection technique filters correlated variables, biases and unwanted noise from the dataset. It selects important features that may improve the model performance. In the present work, RRF model [29] is used for feature selection task. The regularized random forest (RRF) uses one ensemble instead of multiple ensembles and a feature with the highest regularized information gain is inserted at a node based on the instances only at that node. If more than one feature has same regularized information gain, then one of these features is selected arbitrary. RRF model is implemented in R, python and to do tedious task in an easy way. It gives node impurity measured by the Gini index.

According to the RRF algorithm features $F_i$, $F_j$, $F_l$, $F_m$, $F_n$, $F_o$ and $F_x$ are least important for IgG dataset. When target is changed then it will effect the subset of important features. The important features for both the datasets are different. For IgA dataset features $F_i$, $F_j$, $F_k$, $F_l$, $F_m$, $F_n$ and $F_o$ are less important. Table 5 shows the ranking of features based on node purity i.e. computed by RRF. Higher the node purity, higher the rank is. Seven least ranked (24 to 30) features aren't considered for training the model. The feature selection affects the model performance as shown in Table 7.

### *2.4. Machine learning methods*

To get the better results, parameters of models are need to be tuned. The models used in present study is described in Table 6 with required packages and their tuning parameters.

### *2.5. Benchmark of proposed model correctness*

For the benchmarking of model correctness, the performance of proposed model is compared with existing IgPred [25] model. The proposed model is based on seven different models and IgPred is based on single model i.e. SVM. The independent dataset of 44 IgG, 44 IgA and 44 non-IgG, 44 non-IgA epitopes are collected from immune epitope database (IEDB). The epitopes in independent dataset are unique and they are not present in training of IgPred and proposed model. IgPred

**Table 5**
Feature Importance for IgG and IgA epitopes.

| Rank | Features | IncNodePurity$_{IgG}$ | Features | IncNodePurity$_{IgA}$ |
|---|---|---|---|---|
| 1 | $F_{ab}$ | 177.33 | $F_d$ | 11.19 |
| 2 | $F_e$ | 160.75 | $F_{aa}$ | 7.79 |
| 3 | $F_c$ | 155.96 | $F_{ab}$ | 7.49 |
| 4 | $F_d$ | 153.55 | $F_e$ | 6.74 |
| 5 | SL | 143.7 | $F_c$ | 6.41 |
| 6 | $F_z$ | 133.5 | $F_q$ | 6.11 |
| 7 | $F_b$ | 133.33 | SL | 5.94 |
| 8 | $F_{aa}$ | 132.63 | $F_v$ | 5.81 |
| 9 | $F_y$ | 117.86 | $F_p$ | 5.53 |
| 10 | $F_f$ | 112.47 | $F_y$ | 5.31 |
| 11 | $F_a$ | 100.2 | $F_b$ | 5.00 |
| 12 | $F_{ac}$ | 97.5 | $F_z$ | 4.65 |
| 13 | $F_q$ | 71.98 | $F_a$ | 4.60 |
| 14 | $F_p$ | 68.15 | $F_h$ | 4.47 |
| 15 | $F_r$ | 63.06 | $F_f$ | 4.11 |
| 16 | $F_h$ | 56.14 | $F_{ac}$ | 4.03 |
| 17 | $F_v$ | 52.89 | $F_x$ | 3.97 |
| 18 | $F_s$ | 51.95 | $F_t$ | 3.87 |
| 19 | $F_u$ | 50.66 | $F_w$ | 3.80 |
| 20 | $F_t$ | 50.58 | $F_r$ | 3.56 |
| 21 | $F_k$ | 48.63 | $F_s$ | 3.46 |
| 22 | $F_w$ | 44.13 | $F_u$ | 3.43 |
| 23 | $F_g$ | 44.11 | $F_g$ | 3.12 |
| 24 | $F_x$ | 41.17 | $F_l$ | 2.79 |
| 25 | $F_i$ | 36.19 | $F_k$ | 2.62 |
| 26 | $F_l$ | 34.13 | $F_m$ | 2.42 |
| 27 | $F_m$ | 29.67 | $F_i$ | 1.89 |
| 28 | $F_j$ | 27.8 | $F_j$ | 1.44 |
| 29 | $F_n$ | 21.61 | $F_o$ | 1.34 |
| 30 | $F_o$ | 21 | $F_n$ | 1.22 |

prediction on independent dataset is recorded by using its web server (http://crdd.osdd.net/raghava/IgPred/). Independent dataset of epitopes are provided in Table 13 with predictions from IgPred and proposed model. The output of IgPred and proposed model on these epitopes are used for validation on various parameters like gini, AUC, accuracy, MCC, specificity and sensitivity as described in Table 12. The results shows that the proposed model is outperforming as compare to existing model.

## 3. Methodology

The methodology is represented in Fig 1 . At the beginning, peptide sequences are collected from http://crdd.osdd.net/raghava/IgPred/. Dataset contains peptide sequences, mixture of negative and positive epitopes inducing IgA and IgG antibodies with variable length from 4 to 50. The feature measurement is done in second step and explained in Section 2.2. In third step, regularized random forest (RRF) [29] is used to get the subset of important features. This process reduces the space complexity, time complexity as well as increases the accuracy of model. In the fourth step, the dataset is used to train the classifiers, with their optimum tuning parameters. The used machine learning models are presented in Table 6. The models are combined to get proposed multilevel ensemble model as mentioned in Section 3.2. The flow of proposed scheme is presented in Fig 2 . The proposed

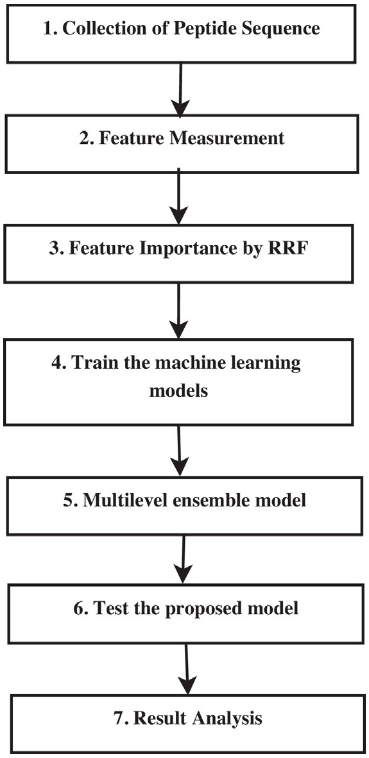
1. Collection of Peptide Sequence
2. Feature Measurement
3. Feature Importance by RRF
4. Train the machine learning models
5. Multilevel ensemble model
6. Test the proposed model
7. Result Analysis

**Fig. 1.** Methodology of proposed model.

**Table 7**
Impact of features on accuracy for IgG and IgA epitopes.

| Number of features | Features | Accuracy$_{IgG}$ | Features | Accuracy$_{IgA}$ |
|---|---|---|---|---|
| 10 | $F_{ab}$ - $F_f$ | 96.6 | $F_d$ - $F_y$ | 98.5 |
| 11 | $F_{ab}$ - $F_a$ | 96.6 | $F_d$ - $F_b$ | 98.4 |
| 12 | $F_{ab}$ - $F_{ac}$ | 96.1 | $F_d$ - $F_z$ | 98.2 |
| 13 | $F_{ab}$ - $F_q$ | 95.5 | $F_d$ - $F_a$ | 98.5 |
| 14 | $F_{ab}$ - $F_p$ | 95.6 | $F_d$ - $F_h$ | 97.9 |
| 15 | $F_{ab}$ - $F_r$ | 95.0 | $F_d$ - $F_f$ | 97.2 |
| 16 | $F_{ab}$ - $F_h$ | 94.9 | $F_d$ - $F_{ac}$ | 97.5 |
| 17 | $F_{ab}$ - $F_v$ | 94.6 | $F_d$ - $F_x$ | 97.5 |
| 18 | $F_{ab}$ - $F_s$ | 93.0 | $F_d$ - $F_t$ | 97.5 |
| 19 | $F_{ab}$ - $F_u$ | 88.1 | $F_d$ - $F_w$ | 97.5 |
| 20 | $F_{ab}$ - $F_t$ | 90.2 | $F_d$ - $F_r$ | 97.5 |
| 21 | $F_{ab}$ - $F_k$ | 91.8 | $F_d$ - $F_s$ | 97.5 |
| 22 | $F_{ab}$ - $F_w$ | 94.4 | $F_d$ - $F_u$ | 97.5 |
| **23** | $F_{ab}$ - $F_g$ | **94.4** | $F_d$ - $F_g$ | **97.5** |
| 24 | $F_{ab}$ - $F_x$ | 93.5 | $F_d$ - $F_l$ | 96.2 |
| 25 | $F_{ab}$ - $F_i$ | 92.4 | $F_d$ - $F_k$ | 95.2 |
| 26 | $F_{ab}$ - $F_l$ | 84.2 | $F_d$ - $F_m$ | 95.4 |
| 27 | $F_{ab}$ - $F_m$ | 72.3 | $F_d$ - $F_i$ | 94.9 |
| 28 | $F_{ab}$ - $F_j$ | 77.4 | $F_d$ - $F_j$ | 94.6 |
| 29 | $F_{ab}$ - $F_n$ | 73.0 | $F_d$ - $F_o$ | 93.8 |
| 30 | $F_{ab}$ - $F_o$ | 70.2 | $F_d$ - $F_n$ | 93.2 |

**Table 6**
Machine learning models.

| Model | Method | Required package | Tuning parameter |
|---|---|---|---|
| Random Forest (RF) [30] | rf | random forest | mtry=2, ntree=500 |
| Support vector machine (SVM) [31] | ksvm | kernlab | kernel=”rbfdot”, type=”C-svc” |
| Decision Tree [32] | rpart | None | usesurrogate=0, maxsurrogate=0 |
| Neural Network [33] | nnet | nnet | size=10 |
| Extreme Learning Machine (ELM) [34] | elmtrain | elmNN | nhid=10 |
| Avnnet [35] | avNNet | caret | size, linout, trace |
| Regularized Random Forest (RRF) [29] | RRF | RRF | None |

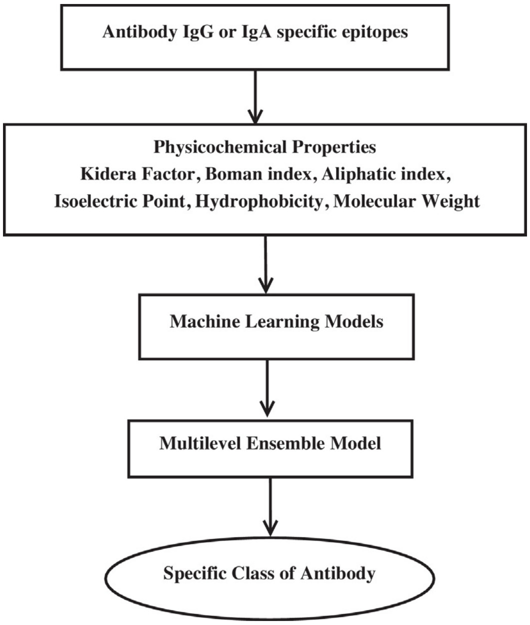
Antibody IgG or IgA specific epitopes
Physicochemical Properties
Kidera Factor, Boman index, Aliphatic index,
Isoelectric Point, Hydrophobicity, Molecular Weight
Machine Learning Models
Multilevel Ensemble Model
Specific Class of Antibody

**Fig. 2.** Flow of proposed scheme.

model is described in Fig 3 . Finally, performance of the model is evaluated on various parameters such as area under the curve (AUC), specificity, sensitivity, gini and accuracy with repeated k-fold cross validation.

### 3.1. *Flow of proposed scheme*

Fig. 2 describes the prediction of IgG, IgA epitopes and their physicochemical properties which is mentioned in Section 2 are used to train the machine learning models. To get proposed model seven models are combined, details are described in Section 3.2. The final prediction is produced by proposed model. The proposed model is separately trained for IgA and IgG antibodies.

### 3.2. *Proposed multilevel ensemble model*

Ensemble is used to deal with the worst case of model prediction. In the present work, focus is on the false prediction as well as true prediction of the model and multilevel ensemble model is used to deal with false and true predictions. Seven models i.e. Decision tree, RF, SVM, ELM, Neural network, RRF and avNNet are combined to get better accuracy as mentioned in Fig 3 . All the models are trained on 70% of the dataset and 30% is used for testing. The proposed model is divided into three phases and all phases are explained below:

**Phase I:** The decision tree, ELM, neural network, SVM model are trained with 70% of dataset and generate predictions from 30% of dataset.

**Phase II:** The false predictions of two models (decision tree and ELM) from phase I are used to train the RF model. The false predictions of two models (neural network and SVM) from phase I are used to train the avNNet model.

**Phase III:** The false predictions from phase II and true predictions from Phase I are combined. This combined new dataset is used to train the RRF model that provides final predictions.

In this approach, true predictions as well as false predictions are refined to get accurate proposed model. The purpose of using true prediction as the input of other models is to deal with false positive results (Non-antigenic is considered as antigenic). The data is travelled through seven models because of this models perfectly learn the data to provide reliable and accurate results.

## 4. Model evaluation

Various parameters such as gini, accuracy, AUC, specificity and sensitivity are calculated to evaluate the performance of model. Repeated k-fold cross validation is performed to test the robustness of proposed model. According to RRF algorithm, 7 features are discarded from both the datasets (IgA and IgG) as mentioned in Section 2.3. Based on rest of the features, formula for IgG antibody prediction is formulated which consist of important features and target class as mentioned below:

$$CL_{IgG} \sim f(F_a, F_b, F_c, F_d, F_e, F_f, F_g, F_h, F_k, F_p, F_q, F_r, F_s, F_t, F_u, F_v, F_w, F_y, F_z, F_{aa}, F_{ab}, F_{ac}) \tag{1}$$

For IgA antibody prediction formula is given below:

$$CL_{IgA} \sim f(F_a, F_b, F_c, F_d, F_e, F_f, F_g, F_h, F_p, F_q, F_r, F_s, F_t, F_u, F_v, F_w, F_x, F_y, F_z, F_{aa}, F_{ab}, F_{ac}) \tag{2}$$

### 4.1. *Performance evaluation*

There are various parameters like gini, accuracy, AUC, specificity and sensitivity to measure the performance of models. In the present study, all these parameters are used for evaluation.

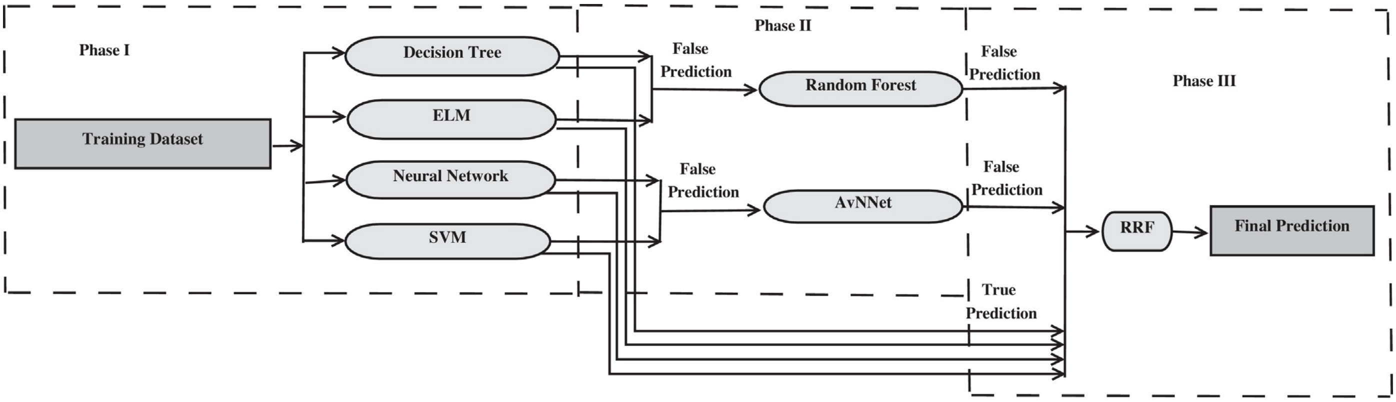

**Fig. 3.** Multilevel ensemble model.

**Table 8**
Performance evaluation of machine learning models for fixed and variable length of IgG epitopes.

| | | Fixed length of epitopes | | | | | Variable length of epitopes | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **SN** | **Model Name** | **Gini** | **ACC** | **AUC** | **Spec** | **Sens** | **Gini** | **ACC** | **AUC** | **Spec** | **Sens** |
| 1 | **Random Forest** | 0.16 | 58.45 | 0.58 | 0.51 | 0.51 | 0.31 | 65.3 | 0.65 | 0.56 | 0.56 |
| 2 | **AvNNet** | 0.06 | 51.47 | 0.53 | 0.50 | 0.45 | 0.21 | 60.6 | 0.61 | 0.52 | 0.54 |
| 3 | **Decision Tree** | 0.16 | 49.97 | 0.42 | 0.40 | 0.46 | 0.23 | 61.1 | 0.62 | 0.51 | 0.56 |
| 4 | **RRF** | 0.19 | 59.84 | 0.59 | 0.52 | 0.52 | 0.31 | 65.8 | 0.66 | 0.56 | 0.56 |
| 5 | **Neural Network** | 0.08 | 54.20 | 0.54 | 0.48 | 0.48 | 0.21 | 60.6 | 0.61 | 0.53 | 0.53 |
| 6 | **ELM** | 0.02 | 49.43 | 0.51 | 0.45 | 0.46 | 0.19 | 50.8 | 0.59 | 0.43 | 0.61 |
| 7 | **SVM** | 0.08 | 54.25 | 0.54 | 0.48 | 0.47 | 0.28 | 63.9 | 0.64 | 0.54 | 0.56 |
| **8** | **Proposed model** | **0.19** | **67.02** | **0.59** | **0.53** | **0.50** | **0.86** | **94.43** | **0.93** | **0.99** | **0.98** |

**Table 9**
Performance evaluation of machine learning models for fixed and variable length of IgA epitopes.

| | | Fixed length of epitopes | | | | | Variable length of epitopes | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **SN** | **Model Name** | **Gini** | **ACC** | **AUC** | **Spec** | **Sens** | **Gini** | **ACC** | **AUC** | **Spec** | **Sens** |
| 1 | **Random Forest** | 0.32 | 63.35 | 0.66 | 0.68 | 0.43 | 0.38 | 69.11 | 0.69 | 0.59 | 0.57 |
| 2 | **AvNNet** | 0.34 | 44.1 | 0.67 | 0.71 | 0.40 | 0.31 | 65.85 | 0.66 | 0.59 | 0.53 |
| 3 | **Decision Tree** | 0.23 | 60.87 | 0.61 | 0.61 | 0.45 | 0.28 | 64.23 | 0.64 | 0.57 | 0.53 |
| 4 | **RRF** | 0.30 | 62.73 | 0.65 | 0.66 | 0.44 | 0.39 | 69.51 | 0.69 | 0.59 | 0.58 |
| 5 | **Neural Network** | 0.14 | 57.14 | 0.57 | 0.56 | 0.44 | 0.34 | 67.07 | 0.67 | 0.58 | 0.55 |
| 6 | **ELM** | 0.05 | 44.1 | 0.52 | 0.44 | 0.50 | 0.19 | 58.94 | 0.59 | 0.51 | 0.53 |
| 7 | **SVM** | 0.33 | 61.49 | 0.66 | 0.70 | 0.41 | 0.36 | 68.29 | 0.68 | 0.59 | 0.55 |
| **8** | **Proposed model** | **0.34** | **61.40** | **0.67** | **0.57** | **0.41** | **0.91** | **97.56** | **0.95** | **0.97** | **0.99** |

#### 4.1.1. *Gini coefficient*

Inequality in the distribution is measured through gini coefficient. The range of gini value is between 0 and 1. Like, model M has gini value 60% and model D has gini value 45% then Model M is considered as an efficient model as compare to model D.

#### 4.1.2. *AUC*

Area under the curve (AUC) is calculated to measure the quality of classifier. The amount of area under the receiver operating characteristics (ROC) curve is AUC. The model scoring high AUC as compared to other models is considered as efficient model. Its value is between 0 and 1. The quality of model is good if it has AUC value near to 1.

#### 4.1.3. *Accuracy*

Accuracy is calculated to measure the correctness of classifier. The accuracy can be calculated as:

$$Accuracy = \frac{TP + TN}{TotalData} * 100 \qquad (3)$$

#### 4.1.4. *Sensitivity*

Sensitivity(Sens) is also known as recall or true positive rate. It is the proportion of actual positives which are correctly identified as positives by the classifier and is computed as:

$$Sensitivity = \frac{TP}{TP + FN} \qquad (4)$$

#### 4.1.5. *Specificity*

Specificity(Spec) is also known as true negative Rate. It relates to the classifiers ability to identify negative results and is computed as:

$$Specificity = \frac{TN}{TN + FP} \qquad (5)$$

TN: True negative, FP: False positive, TP: True positive and FN: False negative

### 4.2. *Repeated K-fold cross validation*

The large number of comparisons are always preferred, to comparison the performance of model. To run K-fold cross validation multiple time or increase the number of comparisons, repeated K-fold cross validation is useful. In K-fold cross validation, only k comparisons are acquired. In cross validation, in each fold random data is provided to do the comparisons. Here, 10-fold cross validation is repeated for 3 times.

## 5. Result analysis, comparison and discussion

Epitope length is important in training the model and it is efficient to use variable length of epitopes. The dataset of IgG and IgA fixed length epitopes are accessible from IgPred [25] in which the number of IgG are 6116 and non-IgG are 6116; also the number of IgA are 267 and non-IgA are 267. The machine learning models are trained with fixed length of IgG, IgA epitopes and evaluated on various parameters as mentioned in Table 8 and Table 9. On all the evaluation parameters, model trained with variable length of epitopes have better results as compare to models trained with fixed length of epitopes.

The models may get biased while training, to handle this issue SMOTE algorithm can be used. Another problem is overfitting/underfitting, to deal with overfitting/underfitting issue, the model should be cross validated and tested on independent dataset, if performance is found to be consistent then models are free from overfitting/underfitting. Overfitting is when model learn too much and underfitting is when model learn too less. In cross validation, models are executed n times and accuracy is recorded if accuracy is highly fluctuating then that model is overfitted/underfitted/biased. In present work, repeated K fold cross validation is used that describes the consistency in the accuracy which means proposed model is not effected from these problems. For validation of proposed model, benchmark dataset is used and compare with the existing model by using various parameters such as gini, AUC, accuracy, MCC, specificity and sensitivity as described in Table 12. The result concludes two things about proposed model. First, the proposed model is free from overfitted/underfitted/biased issues.

**Table 10**
Repeated 10-fold cross validation of IgG and IgA proposed model.

| | IgG Proposed Model | | | IgA Proposed Model | | |
|---|---|---|---|---|---|---|
| **Folds** | **Run 1** | **Run 2** | **Run 3** | **Run 1** | **Run 2** | **Run 3** |
| 1 | 93.10 | 92.41 | 93.36 | 93.17 | 95.61 | 96.10 |
| 2 | 93.07 | 93.48 | 93.22 | 94.15 | 96.10 | 96.10 |
| 3 | 93.56 | 93.02 | 92.99 | 92.20 | 97.07 | 95.12 |
| 4 | 92.93 | 93.62 | 93.28 | 95.12 | 93.66 | 96.10 |
| 5 | 92.41 | 93.36 | 93.19 | 94.15 | 95.12 | 94.63 |
| 6 | 92.41 | 92.24 | 93.13 | 93.66 | 96.10 | 97.07 |
| 7 | 93.82 | 93.05 | 93.53 | 92.68 | 97.56 | 92.20 |
| 8 | 93.79 | 92.76 | 92.39 | 94.63 | 90.73 | 94.63 |
| 9 | 93.91 | 93.45 | 93.45 | 95.12 | 97.07 | 93.66 |
| 10 | 93.25 | 93.05 | 93.51 | 92.68 | 93.66 | 94.63 |

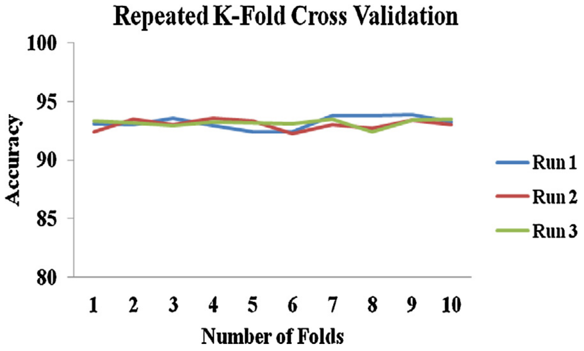

**Fig. 4.** Repeated K-fold cross validation of IgG proposed model.

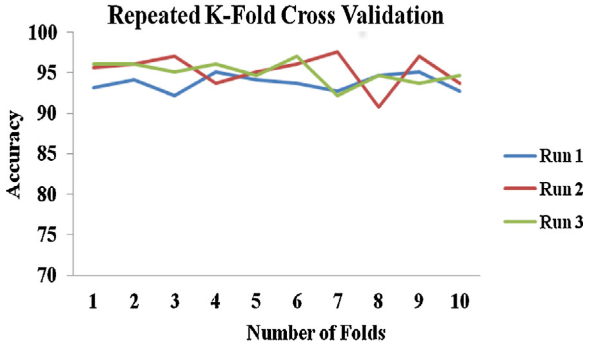

**Fig. 5.** Repeated K-fold cross validation of IgA proposed model.

**Table 11**
Performance comparison with existing model and the proposed model.

| | IgG Epitopes | | IgA epitopes | |
|---|---|---|---|---|
| **Parameters** | **IgPred** | **Proposed model** | **IgPred** | **Proposed model** |
| **Acc(%)** | 70.42 | **94.43** | 72.07 | **97.56** |
| **MCC** | 0.41 | **0.86** | 0.44 | **0.89** |
| **AUC** | 0.76 | **0.93** | 0.78 | **0.95** |

MCC: Matthews correlation coefficient

**Table 12**
Performance comparison on benchmark dataset with existing model and the proposed model.

| | IgG epitopes | | IgA epitopes | |
|---|---|---|---|---|
| **Parameters** | **IgPred** | **Proposed model** | **IgPred** | **Proposed model** |
| **Acc(%)** | 73.86 | 86.36 | 54.55 | 77.27 |
| **AUC** | 0.77 | 0.86 | 0.57 | 0.77 |
| **Gini** | 0.54 | 0.72 | 0.14 | 0.54 |
| **Spec** | 0.86 | 0.90 | 0.62 | 0.74 |
| **Sens** | 0.67 | 0.83 | 0.52 | 0.81 |
| **MCC** | 0.50 | 0.73 | 0.11 | 0.55 |

Second, the outcome of proposed model is improved as compare to the existing technique.

Table 7 depicts the accuracy on various subset of features. The seven least ranked features are discarded because they are affecting the accuracy of the classifier. To balance the number of features and accuracy, 23 feature ($F_{ab}$ - $F_g$ in IgG antibody and $F_d$ - $F_g$ in IgA antibody) are considered to train the model.

Table 6 describes the machine learning models that are trained on the dataset with optimum tuning parameters. The dataset is partitioned into two parts 70% and 30%. The prepared models are unknown to the 30% of dataset. The proposed model is combination of seven models that makes it a multilevel ensemble model as discussed in Section 3.2. The models are evaluated on various parameters as mentioned in Table 8 and Table 9. From the results, it is concluded that the accuracy of proposed model is increased as compared to the single model accuracy.

Table 10 describes the accuracy of the proposed model. The accuracy has been recorded by applying 10-fold cross validation 3 times. For cross validation, 70% of dataset is used for training and 30% used for testing. Figs. 4 and 5 describe the accuracy of proposed model 3 times in 10 runs and shows the consistency in the accuracy of proposed model.

### 5.1. Performance comparison

The dataset is taken from IgPred [25]. The IgA, IgG incuding epitopes and non-IgA, non-IgG epitopes of length 4 to 50 are available on http://crdd.osdd.net/raghava/IgPred/ which are used to train the proposed model. In the present work, balanced data is used. IgPred uses features like binary profiles, dipeptide composition and amino acid composition. The outcome of dipeptide composition based model is better than other models in IgPred. In the present study, results of dipeptide composition based model is compared with the proposed model. The results recommend that the proposed model outperforms as contrast with the current model. The Table 11 describes the performance of IgPred and proposed model (IgG, IgA).

To validate, the existing model and proposed model are evaluated on various parameters such as gini, AUC, accuracy, MCC, specificity and sensitivity, benchmark dataset is used as mentioned in Section 2.5. These epitopes prediction is calculated from IgPred server and compared with outcome of proposed model as shown in Table 12. The result shows that the proposed model is outperforming as contrast to existing model.

## 6. Conclusion

The proposed model increases the prediction accuracy of IgG and IgA antibodies as compared to the existing technique. In the present study, seven models i.e. decision tree, ELM, RF, neural netwrok, SVM, avnnet and RRF are used to create multilevel ensemble model. A novel multilevel ensemble model is developed for prediction and it produces high accuracy, gini, AUC, specificity and sensitivity with variable length of epitopes. The multilevel ensemble model is divided into 3 phases. In this approach, true predictions and false predictions are used to get accurate proposed model. The benefit of using true prediction as the input of other models is to deal with false positive results. The data is travelled through seven models because of this models perfectly learn the data to provide reliable and accurate results. The proposed model is compared with existing IgPred model and validated on benchmark dataset. To check the robustness of proposed model repeated k-fold cross validation is used. The modelled peptide sequences used in present study are accessible from http://crdd.osdd.net/raghava/IgPred/.

**Table 13**
Benchmark dataset of IgG and IgA epitopes.

| Sr. No. | *IgG _S equence* | Actual | Proposed Model | IgPred Model | *IgA _S equence* | Actual | Proposed Model | IgPred Model |
|---|---|---|---|---|---|---|---|---|
| 1 | AATGAATAAA | 1 | 1 | 0 | AEVLKDAIKDLVMTKPAPTC | 0 | 0 | 1 |
| 2 | AEFYLNPDQSGEFMFDFDGDEIF | 1 | 1 | 1 | AGKREMVIIT | 1 | 1 | 1 |
| 3 | AFYGVWPLL | 0 | 1 | 0 | AHGRSQVLQQSTYQLLQELCC | 0 | 0 | 1 |
| 4 | AILDMIAGAHWGVLAGIA | 0 | 0 | 0 | AQGSVQPQQLPQFEEIRNLAL | 0 | 0 | 1 |
| 5 | AKAPPAPKPAPQPGP | 1 | 1 | 1 | ASLKMADPNRFR | 1 | 1 | 1 |
| 6 | AKQELE | 1 | 1 | 1 | ASREAK | 1 | 1 | 1 |
| 7 | ATRKTSER | 0 | 0 | 0 | ASREAKKQVEKALE | 1 | 1 | 0 |
| 8 | AYALYGVWPL | 0 | 0 | 0 | DAEFRHDSGYEVHHQKLVFFAED VGSNKGAIIGLMVGGVVIA | 1 | 0 | 0 |
| 9 | CFHQGKEYAPG | 1 | 1 | 1 | DEKGIMRTGLISFENNNYYFNENGE | 1 | 1 | 1 |
| 10 | CITQYERESQAYY | 1 | 1 | 1 | EQSRCQAIHN | 1 | 1 | 1 |
| 11 | CNRLGGLFNFGPKQKI | 1 | 1 | 1 | ERMKDTLRIT | 1 | 1 | 1 |
| 12 | DLEDRDRSELSPLLLTTT | 0 | 0 | 1 | ESMAGKREMV | 1 | 0 | 1 |
| 13 | DLLVGSATLCSALYVGDL | 0 | 0 | 0 | FFQP | 0 | 1 | 0 |
| 14 | DPNAVCETDKWKYENPCKKM | 1 | 1 | 1 | FPPQLPYPQP | 0 | 0 | 1 |
| 15 | DVKFPGGG | 0 | 0 | 1 | FTTKVIGKDSRDFDISPKV | 1 | 1 | 1 |
| 16 | EGHLIDLKRV | 1 | 1 | 1 | GGSGGRGRGGSGGRGRGGSG | 0 | 0 | 1 |
| 17 | EKPHFP | 1 | 1 | 1 | GIAKAFRAPSIREVSPG FGTLTQGGASIMYGN | 1 | 0 | 0 |
| 18 | EPWFLHGLGLARTYWRDTNTG | 1 | 1 | 1 | GKREMVIITF | 1 | 0 | 1 |
| 19 | ESDEAPFMFSENKFL | 1 | 0 | 1 | GSENKRTGALGNLKN | 1 | 1 | 1 |
| 20 | EVNQIVETNVRLRQQW | 1 | 1 | 1 | GSPPRRPPPGRRPFFHPVGE | 0 | 0 | 0 |
| 21 | FKNPHAKKQDVV | 1 | 1 | 1 | IDKLCVWNNK | 1 | 1 | 1 |
| 22 | FSPRRHWTTQGCNCSIYP | 0 | 0 | 0 | IILHQQHH | 0 | 1 | 0 |
| 23 | GCNCSIYPGHITGHRMAW | 0 | 1 | 0 | IPEQ | 0 | 0 | 0 |
| 24 | GGQIVGGV | 0 | 0 | 1 | ISIKYDPRKDSEVFA | 1 | 1 | 1 |
| 25 | GIGAVLKVLTTGLPALISWIKR | 1 | 1 | 1 | ITFKSGATFQ | 1 | 0 | 1 |
| 26 | GNASRCWVAMTPTVATRD | 0 | 0 | 0 | ITIMDNGNIDTELLVGTLTLGGY | 1 | 1 | 1 |
| 27 | GSAGHTVSGFVSLLAPGA | 0 | 1 | 1 | KASITEIKADKT | 1 | 1 | 1 |
| 28 | GVDAETHVTGGSAGHTVS | 0 | 0 | 1 | KGINLIDDIKYYFDEKGIMRTGLIS | 1 | 0 | 1 |
| 29 | GVRATRKT | 0 | 0 | 1 | KREMVIITFK | 1 | 0 | 1 |
| 30 | GYGAGVAGAL | 0 | 0 | 0 | LALLAIVATTATTAVRVPVPQ | 0 | 1 | 0 |
| 31 | HADPAPASAENVKEIIELLKGL DLRLQTVEGKVDKILA | 1 | 1 | 1 | LCCQHLWQIPEQSQCQAIHNV | 0 | 0 | 1 |
| 32 | HYAPRPCGI | 0 | 0 | 0 | LCVWNNKTPN | 1 | 1 | 1 |
| 33 | HYKLFLARL | 0 | 0 | 0 | LQQQLIPCRD | 1 | 1 | 1 |
| 34 | KNQVEGEVQIVSTATQTFLA | 0 | 0 | 0 | LRITYLTETK | 1 | 0 | 1 |
| 35 | KSGNFKHLREFVFK NKDGFLYVYKGYQPIDV | 1 | 1 | 1 | LRLQTAGNVDHVGLGT | 1 | 1 | 1 |
| 36 | KTNTPADVFIVFTDNETFAG | 1 | 1 | 1 | MKTFLILALLAIVATTATTAV | 0 | 0 | 0 |
| 37 | KTSERSQP | 0 | 0 | 1 | MSDGAVQPDGGQPAVRNERATG | 1 | 1 | 0 |
| 38 | LEGAARQ | 1 | 1 | 1 | MVIITFKSGA | 1 | 1 | 1 |
| 39 | LGVRATRK | 0 | 0 | 1 | PGEGPSTGPRGQGDGGRRKK | 0 | 1 | 1 |
| 40 | LPATQLRRHIDLLVGSAT | 0 | 0 | 0 | PLKP | 1 | 1 | 1 |
| 41 | MMNWSPTTALVMAQLLRI | 0 | 1 | 1 | PLLGCIGSTCAEDGN | 1 | 1 | 1 |
| 42 | MYVGGVEHRL | 0 | 0 | 1 | PPDQLVNLHDFRSD EIEHLVVEE | 1 | 1 | 0 |
| 43 | NPDPNPN | 1 | 1 | 1 | PQQPISQQQQQQQQQQQQQQQ | 0 | 0 | 1 |
| 44 | NQVYYRPMDEYSN | 1 | 1 | 1 | QEQKQQLQQQ | 0 | 0 | 0 |
| 45 | NRRPQDVK | 0 | 0 | 1 | QFLGQQQPFPPQQPYPQPQPF | 0 | 0 | 1 |
| 46 | NSTNSGIN | 1 | 1 | 1 | QGSFRPSQQNPQAQGSVQPQQ | 0 | 0 | 1 |
| 47 | NTHVTGAVQGHGAF TLTSLFQPGASQKIQLV | 1 | 1 | 1 | QLVKDKNIDISIKYDPRKDSE | 1 | 0 | 1 |
| 48 | PGGGQIVG | 0 | 0 | 0 | QPFPPQLPYP | 0 | 0 | 1 |
| 49 | PIPKARRP | 0 | 0 | 0 | QPFPQPQLPYSQPQPFRPQQP | 0 | 0 | 1 |
| 50 | PLATQPPVLAL | 1 | 1 | 1 | QPFPSQQPYLQLQPFPQPQLP | 0 | 0 | 1 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 51 | PPAYEK | 1 | 1 | 1 | QPQEQVPLVQQQQFLGQQQPF | 0 | 0 | 1 |
| 52 | PPGEFLQVSIQDTRNAVRAC | 1 | 1 | 1 | QPQYSQPQQPISQQQQQQQQQ | 0 | 0 | 1 |
| 53 | PQDVKFPG | 0 | 0 | 1 | QPYLQLQPFPQPQLPYSQPQP | 0 | 0 | 1 |
| 54 | PRRGPRLGVRAPRKTS | 0 | 1 | 1 | QPYPQPQPFPSQQPYLQLQPF | 0 | 0 | 1 |
| 55 | QDVKFPGG | 0 | 0 | 1 | QQLQQQQQQQ | 0 | 0 | 1 |
| 56 | QDVKFPGGGQIVGG VYLLPRRGPRL | 0 | 0 | 1 | QQQQLQQQQQ | 0 | 0 | 1 |
| 57 | QKTRTSRRAKPPQRPKQQPAAP | 1 | 1 | 1 | QQQQQQQQQQQQILQQILQQ | 0 | 0 | 1 |
| 58 | QPIPKARR | 0 | 0 | 0 | QQQQQQQQQQQQQQQQQQQIL | 0 | 0 | 1 |
| 59 | RATRKTSE | 0 | 0 | 0 | QQYPLGQGSFRPSQQNPQAQG | 0 | 0 | 1 |
| 60 | REQAPNLVY | 1 | 1 | 1 | REMVIITFKS | 1 | 0 | 1 |
| 61 | RGQRTKTNARTRKGPRKPIKK | 1 | 1 | 1 | RGRGRGEKRPRSPSSQSSSS | 0 | 1 | 0 |
| 62 | RKTSERSQ | 0 | 0 | 1 | RRPFFHPVGEADYFEYHQEG | 0 | 0 | 0 |
| 63 | RLGVRATR | 0 | 0 | 1 | SATAIFDTTTLNPTIAGAGDVKASAE GQLG | 1 | 1 | 1 |
| 64 | RPEPKKPWSGVWNASTY | 1 | 1 | 1 | SEEND | 1 | 1 | 1 |
| 65 | RSQPRGRR | 0 | 0 | 0 | SELLSLINDMPITNDQKKLMSNNV | 1 | 0 | 1 |
| 66 | SAFVFPTKD | 1 | 0 | 1 | SFQQPLQQYPLGQGSFRPSQQ | 0 | 0 | 1 |
| 67 | SAQSGTSGTSAQSGT | 1 | 0 | 1 | SGPRHRDGVRRPQKRPSCIG | 0 | 1 | 1 |
| 68 | SCLTVPASAYQVRNSTGL | 0 | 1 | 0 | SLLTEVETPIRNEWGCRCNDSSD | 1 | 1 | 0 |
| 69 | SMVGNWAKVLVVLLLFAG | 0 | 1 | 0 | SMAGKREMVI | 1 | 0 | 1 |
| 70 | SQWN | 1 | 1 | 1 | SQQNPQAQGSVQPQQLPQFEE | 0 | 0 | 1 |
| 71 | SSKYGDTSTNNVRG DLQVLAQKAERTLP | 1 | 1 | 1 | SQVLQESTYQ | 0 | 0 | 1 |
| 72 | TCSMFVYGGC | 1 | 1 | 1 | STYQLVQQLC | 0 | 1 | 1 |
| 73 | TISSLQS | 1 | 1 | 0 | TAVRVPVPQLQPQNPSQQQPQ | 0 | 0 | 1 |
| 74 | TPGCVPCVREGNASRCWV | 0 | 1 | 0 | TETKIDKLCV | 1 | 1 | 1 |
| 75 | TRKTSERS | 0 | 0 | 0 | TFKSGATFQV | 1 | 1 | 1 |
| 76 | TSGTSGTSGTSPSSR | 1 | 0 | 1 | TINKPKGYVGKE | 1 | 0 | 1 |
| 77 | TWGENETDVLLLNNTRPPQ | 1 | 1 | 1 | TLRITYLTET | 1 | 1 | 1 |
| 78 | VDPLPSGYQFNPEATKAASP | 1 | 1 | 1 | TRLSRTIGYTVK | 1 | 1 | 1 |
| 79 | VENGLISRVLDGLV | 1 | 1 | 0 | TSQDGNNHQFT | 1 | 1 | 1 |
| 80 | VFVGLILLTL | 0 | 0 | 0 | VATTATTAVRVPVPQLQPQNP | 0 | 0 | 1 |
| 81 | VKEFLESSPNTQWELRAFMA | 1 | 1 | 1 | VETEDTKEPGVLMG GQSESVEFTKDTQTGM | 1 | 1 | 1 |
| 82 | VKTIGDKRTLTLNTTANYT | 1 | 1 | 0 | VKAETRLNPDLQPTE | 1 | 1 | 1 |
| 83 | VRATRKTS | 0 | 0 | 1 | VLQQSTYQLLQELCCQHLWQI | 0 | 0 | 1 |
| 84 | VTSVSAVASGHYLHR | 1 | 1 | 1 | VVLQQHNIAH | 0 | 0 | 1 |
| 85 | VYEAADAILHTPGCVPCV | 0 | 0 | 1 | WQIPEQSQCQAIHNVVHAIIL | 0 | 0 | 1 |
| 86 | WGVLAGIAYFSMVGNWAK | 0 | 0 | 0 | YLLPRRGPRL | 0 | 0 | 0 |
| 87 | WHLNSTALNCNDSLNTGW | 0 | 0 | 0 | YPQPQPQYSQ | 0 | 0 | 1 |
| 88 | YWPPPQGRRRF | 1 | 1 | 1 | YQLLQELCCQHLWQIPEQSQC | 0 | 0 | 1 |

1 represents IgG or IgA, 0 represents Non-IgG or Non-IgA.

We believe that by using more physicochemical properties and machine learning models with their optimized parameters produce even better outcomes.

## References

[1] L. Yel, Selective iga deficiency, J. Clin. Immunol. 30 (1) (2010) 10–16.
[2] A.O. Vladutiu, Immunoglobulin d: properties, measurement, and clinical relevance, Clin. Diagnostic Lab. Immunol. 7 (2) (2000) 131–140.
[3] J.L. Fahey, Antibodies and immunoglobulins: I. structure and function, JAMA 194 (1) (1965) 71–74.
[4] J.F. Ludvigsson, M. Neovius, W. Ye, L. Hammarström, Iga deficiency and risk of cancer: a population-based matched cohort study, J. Clin. Immunol. 35 (2) (2015) 182–188.
[5] T. Chin, Iga and igg subclass deficiencies.
[6] A. Schlessinger, Y. Ofran, G. Yachdav, B. Rost, Epitome: database of structure-inferred antigenic epitopes, Nucleic acids research 34 (suppl 1) (2006) D777–D780.
[7] J. Parker, D. Guo, R. Hodges, New hydrophilicity scale derived from high-performance liquid chromatography peptide retention data: correlation of predicted surface residues with antigenicity and x-ray-derived accessible sites, Biochemistry 25 (19) (1986) 5425–5432.
[8] J. Pellequer, E. Westhof, M. Van Regenmortel, Predicting location of continuous epitopes in proteins from their primary structures, Methods Enzymol. 203 (1990) 176–201.
[9] A. Kolaskar, P.C. Tongaonkar, A semi-empirical method for prediction of antigenic determinants on protein antigens, FEBS Lett. 276 (1-2) (1990) 172–174.
[10] J.-L. Pellequer, E. Westhof, M.H. Van Regenmortel, Correlation between the location of antigenic sites and the prediction of turns in proteins, Immunol. Lett. 36 (1) (1993) 83–99.
[11] A.J. Alix, Predictive estimation of protein linear epitopes by using the program people, Vaccine 18 (3) (1999) 311–314.
[12] M. Odorico, J.-L. Pellequer, Bepitope: predicting the location of continuous epitopes and patterns in proteins, J. Mol. Recognit. 16 (1) (2003) 20–22.
[13] S. Saha, G. Raghava, Bcepred: prediction of continuous b-cell epitopes in antigenic sequences using physico-chemical properties (2004) 197–204.
[14] S. Saha, G. Raghava, Prediction of continuous b-cell epitopes in an antigen using recurrent neural network 65 (1) (2006) 40–48.
[15] J. Chen, H. Liu, J. Yang, K.-C. Chou, Prediction of linear b-cell epitopes using amino acid pair antigenicity scale, Amino Acids 33 (3) (2007) 423–428.
[16] Y. EL-Manzalawy, D. Dobbs, V. Honavar, Predicting linear b-cell epitopes using string kernels, J. Mol. Recognit. 21 (4) (2008) 243–255.
[17] B. Yao, L. Zhang, S. Liang, C. Zhang, Svmtrip: a method to predict antigenic epitopes using support vector machine to integrate tri-peptide similarity and propensity, PloS One 7 (9) (2012) e45152.
[18] J.-H. Huang, M. Wen, L.-J. Tang, H.-L. Xie, L. Fu, Y.-Z. Liang, H.-M. Lu, Using random forest to classify linear b-cell epitopes based on amino acid properties and molecular features, Biochimie 103 (2014) 1–6.
[19] L. Yao, Z.C. HUANG, G. Meng, X.M. PAN, An improved method for predicting linear b-cell epitope using deep maxout networks, Biomed. Environ. Sci. 28 (6) (2015) 460–463.
[20] W. Shen, Y. Cao, L. Cha, X. Zhang, X. Ying, W. Zhang, K. Ge, W. Li, L. Zhong, Predicting linear b-cell epitopes using amino acid anchoring pair composition, BioData mining 8 (1) (2015) 1.
[21] S. Saha, G. Raghava, Algpred: prediction of allergenic proteins and mapping of ige epitopes, Nucleic acids research 34 (suppl 2) (2006) W202–W209.
[22] M.B. Stadler, B.M. Stadler, Allergenicity prediction by protein sequence, The FASEB Journal 17 (9) (2003) 1141–1143.
[23] K. Björklund, D. Soeria-Atmadja, A. Zorzet, U. Hammerling, M.G. Gustafsson, Supervised identification of allergen-representative peptides for in silico detection of potentially allergenic proteins, Bioinformatics 21 (1) (2005) 39–50.
[24] H. Mohabatkar, M. Mohammad Beigi, K. Abdolahi, S. Mohsenzadeh, Prediction of allergenic proteins by means of the concept of chou's pseudo amino acid composition and a machine learning approach, Med. Chem. 9 (1) (2013) 133–137.
[25] S. Gupta, H.R. Ansari, A. Gautam, G.P. Raghava, Identification of b-cell epitopes in an antigen for inducing specific class of antibodies, Biol. Direct 8 (1) (2013) 1.
[26] P. R.-V. Daniel Osorio, R. Torres, Calculate indices and theoretical properties of protein sequences. URL https://github.com/dosorio/Peptides/.
[27] H. Boman, Antibacterial peptides: basic facts and emerging concepts, J. Internal Med. 254 (3) (2003) 197–215.
[28] G. F. Heike Hofmann, Eric Hare, Evaluation of diversity in nucleotide libraries. URL https://github.com/heike/peptider.
[29] S. RColorBrewer, H. Deng, M. H. Deng, Package 'rrf'.
[30] A. Liaw, M. Wiener, Classification and regression by randomforest, R news 2 (3) (2002) 18–22.
[31] S.S. Keerthi, E.G. Gilbert, Convergence of a generalized smo algorithm for svm classifier design, Machine Learn. 46 (1-3) (2002) 351–360.
[32] P.D. Berger, A. Gerstenfeld, A.Z. Zeng, How many suppliers are best? a decision-analysis approach, Omega 32 (1) (2004) 9–15.
[33] B. Ripley, W. Venables, M. B. Ripley, Package 'nnet'.
[34] T. Hothorn, P. Buehlmann, T. Kneib, M. Schmid, B. Hofner, F. Sobotka, F. Scheipl, M. B. Hofner, Package 'mboost'.
[35] C. K. Williams, A. Engelhardt, T. Cooper, Z. Mayer, A. Ziem, L. Scrucca, Y. Tang, C. Candan, M. M. Kuhn, Package 'caret'.