International Journal of Research Studies in Computer Science and Engineering (IJRSCSE)
Volume 8, Issue 1, 2021, PP 23-30
ISSN 2349-4840 (Print) & ISSN 2349-4859 (Online)
DOI: http://dx.doi.org/10.20431/2349-4859.0801003
www.arcjournals.org

Predictive Analysis of Linear B-Cell Epitopes in Immune Function using ML Algorithms

Mr. Adithya D A, Dr. Gowrishankar

Department of Computer Science and Engineering, BMS College of Engineering, Bengaluru, India

****Corresponding Author:** **Adithya D A,** Department of Computer Science and Engineering, BMS College of Engineering, Bengaluru, India*

Abstract: *Antibodies have become vital means for biotechnological as well as many clinical applications such as diagnostic test, vaccine based on peptide, disease prevention, antibody production and also treatment. Normally, one can bind the molecular target called antigen by identifying a slice of its structure namely epitope in an extremely explicit way. This capability to foresee epitopes from antigen sequences is very difficult job. Despite great effort, the accuracy of epitope prediction methods has only progressed to a limited extent, particularly for those that rely on the antigen sequence. In this direction, proposed study aims at review and implement ML solutions for classification of linear and nonlinear B-Cell epitopes.*

Keywords: *Machine Learning (ML), Antigen, Epitope, Linear B-Cell, predictive analysis.*

1. Introduction

When our body is attacked by a parasites, bacteria or virus, alarm of an immune system goes off, setting off a chain response of cellular action. Macrophages/other innate immune cells, say dendritic cells, basophils, neutrophils, organized to help attack entering pathogen. These cells are frequently the ones who carry out the tasks, and the intruder is eliminated. As soon as the body requires a more sophisticated attack, it will turns into T-cells in the thymus then bone marrow- or bursa-derived cells (B-cells).The adaptive immune response's major cellular components are these cells. Further, T- cells are involved in cell-mediated immunity, whereas B -cells are responsible of humoral immunity, which comprises antibodies. These cells are the immune system's unique ops, or a line of defense that learns to recognize specific foreign dangers based on previous contacts and behaviors and attacks them when they repeat.

These cells play a dangerous role in cancer progression and treatment. T-cells are the focus of two new immunotherapies: checkpoint inhibitors, which are FDA-approved to treat a variety of cancers, and CAR T-cell therapy, is now being investigated in medical trials as a potential treatment for blood cancers like lymphoma and leukemia. [1].

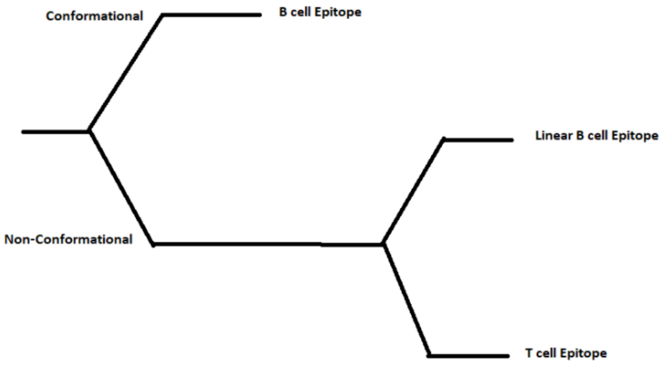

Figure1.*Types of Epitopes*

T-cells and B-cells are referred to as lymphocytes in Figure 1. Lymphocytic development is influenced by primary and secondary lymphoid organs. The bone marrow and the thymus are the key lymphoid tissues in the early development of T- and B-lymphocytes.B-cells compete with viruses and bacteria by producing Y-shaped proteins named "antibodies," which are unique to each pathogen. Antibodies also block entering cell's exterior and label it for harm by other immune cells.B-lymphocytes and cancer have been labelled as having a "hate-love" relation. B-cells, for example, prevent tumor growth by production of antibodies that fight oncogenic viruses (cancer cells), such as the human papillomavirus (HPV), which is accountable for the majority of anal, cervical, and penile growths.Immune-suppressive cytokines are released by regulatory B-cells, which suppress the anti-swellingreaction. B-cells are also far-off prone than T-cells to develop into a fluidgrowth like "Chronic Lymphocytic Leukemia." (CLL) is a type of cancer (B-cell lymphoma) [1].

The region of a antigen that is predicted by the immune system, precisely antibodies, T cells, or B cells, is known as a antigenic determinant .For example, the epitope is the particular portion of the antigen to which an antibody fixes.

Several copies for predicting linear or conformational B-cell epitopes has been established, but their accuracy remains a problem. Several epitope predictions models, including learning-based approaches, have also been created. The model's performance, however, is still not optimal. The primary issue with learning-based prediction models is the imbalance of classes. Because of its unstructured and heterogeneous structure, representing amino acids for machine learning algorithms is difficult.

There are two types of supervision: monitored and unsupervised. ML-based approaches focus on performing data-based predictions and deal with the automatic learning of computers without being explicitly programmed. In the subject of bioinformatics, machine learning has a number of applications. [19]. These methods allows computer programs to perform complex predictions on very large datasets.

Immunologists can use B-cell epitope prediction to create diagnostic tests, peptide-based vaccines, disease prevention, antibody generation, and treatment. T-cell epitope prediction has failed to match the presentation of flexible stretch for B-cell epitope estimation. Thanks to the growing availability of validated epitope databases, bioinformatics professionals can use Machine Learning procedures on curated statistics to design and code enhanced estimation tools for biological scholars. [2].

2. Review of Literature

Since the 1980s, when the opening approach was discovered, researchers have been interested in Linear B-cell epitope prediction. Identification of the B-cell epitope using an precise estimation-based technique can initiate a more efficient and cost-effective serum strategy method. Numerous B-cell epitope prediction approaches has been established over the last two eras. One of the study [3], reviews the present performance and procedure of selected and broadly utilized linear B-cell epitope predictors specifically, BepiPred, ABCpred, BcePred SVMTriP, COBEpro, LBEEP also LBtope. This work aims to address the techniques' performance difficulties by building a consensus classifier that integrates the distinct predictions of these approaches into a one output. A large unbiased data set was utilized to assess the performance of these techniques. While all predictors achieved slightly well than arbitrary categorization against the test data set, these techniques fared worse than indicated in the original sources. The method was compared with necessary cautions, and this improvement in performance can guide investigators in the selection of a predictor while they are doing their investigation.

One of the most essential phases in developing effective vaccines against viruses is identifying epitopes that elicit significant responses from B-cells. Because experimental epitope determination is costly, it necessitates more effort and time. As a result, ML techniques for reliable recognition of B-cell epitopes are crucial. In recent years, different ML algorithms for predicting B-cell epitopes have been developed [4].

Chandra, S. Singh et al [5],the findings reveal that epitopes favour charged plus polar amino acids in general. Epitopes are enhanced by a loop as a supplementary structural feature, which makes them more flexible and reveals a new perspective on the "antibody–antigen interface."

In order to advance the precision of linear B-cell epitope expectation built on the accessibility of empirically recognised linear B-cell epitopes, ML methods are extensively used [6]. BepiPred [7] integrates two amino acid tendency metrics, Parker hydrophilicity and Levitt subordinate construction, with an epitope estimation Hidden Markov Model (HMMOn linear epitopes, Bepipred was effective. However, as compared to systems that depend on training of amino acid physicochemical assets, there was only a minor enhancement in estimate accuracy.

ABCPred [8] used an Artificial Neural Network (ANN) to estimate linear B-cell epitopes. A non-terminated list of 700 B-cell epitopes got from the Bcipep database, as well as 700 non-epitope peptides got from the Swiss Prot list and utilized for estimation. On this record, 5-fold cross authentication trials were used to calculate both recurrent neural networks and feed-forward. The finest enactment, 65.93 percent correctness, achieved by means of a recurrent neural network proficient on peptides of length size 16. Input classification frames varying from 10 to 20 amino acids were verified with the finest enactment, 65.93 percent correctness, stood achieved utilising a recurrent neural network proficient on peptides of length size 16.

Nearest- neighbour and Decision Trees (DT) methods are the two ML methods, tested by Sollner et.al [9]. In this work, they integrated these approaches by feature selection on 1478 characteristics removed after a variability of susceptibility scales, neighbourhood conditions also particular prospect outcomes. The estimation is 72% after verified with a dataset of 1211 B-cell epitopes and 1211 non epitopes by five-fold cross-authentication.

For computing linear B-cell epitopes, Chen projected the Amino Acid Pair (AAP) propensity measure [10]. Bcipep record [11] provided the B-cell epitope data set, which is a group of experimentally determined B-cell epitopes. An AAP antigenicity balance was established, which assigns a propensity outcome to both dipeptides. AAPs are created by the continuous degradation of protein peptides. Cheng's record includes 872 positive epitopes then 872 negative non-epitopes. The AAP antigenicity measure method has an accuracy of 71% when only the AAP tendency measure is employed, according to research utilising the Support Vector Machine classifier. When the AAP measure is integrated with turns, antigenicity, elasticity, hydrophilicity also approachability, the accuracy is 72.5 percent. In this blend approach, the applicable SVM restrictions were 2 = 2 and C equal to 32.

-AAT-fs [12] is a programme developed by L. Wang for estimating rectilinear epitopes based on the antigenicity of the Amino Acid Triplet (AAT). Later, an SVM for the grouping was created using the AAT measure to generate input trajectories. The SVM is capable of operating a Radial Basis Function kernel on homology condensed records by means of fivefold cross validation. AAT-fs technique achieves a 74% higher accuracy in presentation than the AAP measure and other remaining B-cell epitope estimating techniques. With range kernel, incongruity kernel, local arrangement kernel, and sequence kernel, El-Manzalawy et al. presented four kernel roles hooked on SVM. Given the best results, BCPred is the name of the sequence kernel approach [13]. They mentioned a homology condensed record of 701 linear B-cell epitopes mined as of the Bcipep list and 701 non-epitopes mined arbitrarily from SwissProt. BCPred uses a novel type of string kernel-builtSVM to determine 12, 14, 16 and 18also 20-mer long epitopes on or after an arrangement. By the subsequence kernel, the highest accuracy of 75.8% was achieved. EL-Manzalawy also used two alternative strategies to develop stretchable length B-cell epitope estimate models. Kernel functions and deals with the stretchable length epitopes straight, are one method. Four kernel functions, as well as associated methods, were reused for estimating stretchable length epitopes. Mapping stretchable length classifications into stable length feature trajectories is another way. The model established on the subsequence kernel dubbed FBCPred [14] produced the finest results among the other techniques.

Linear B-cell epitopes have a long stretchy length which were identified by W.Zhang and Y.Niu's -BPairwise [15]. Using the Smith Waterman (SW) method, which turns stretchy length peptides into static-length feature trajectories, an encrypting arrangement based on couple wise order is created. The SVM was then utilized to create estimation models as a sorting device. Using this approach, they were able to attain a 66% accuracy.

L. JK. Wee, D.Simarmata's -BayesB [16] is a Bayes Feature Extraction (BFE)-based SVM estimation model for linear B-cell epitopes of various lengths. The range of measurement is 12 to 20, with a precision of 74.50 percent. The -Linear Epitope Prediction System (LEPS) by H.W.Wang, Y.C.Lin, and H.T.Chang is an estimating model [17] based on biological characteristics and SVM. SVM's arithmetical characteristics were effectively employed to record epitope and non-epitope sections of 2, 3and 4 deposits in dimension. In this approach, AntiJen, HIV, and PC records were recycled and well specificity, accuracy, also “Positive Prediction Value” (PPV) were attained in utmost testing instances. For an effective and efficient test method, a linear epitope estimation technique with high specificity and PPV is required. T. Liu et al. applied a feed forward deep neural network [18] to a significant volume of linear B-cell epitope data in the IEDB folder with trial indication and created collaborative estimation models that outperformed current representations.

Table 1 shows the results of a comparison of recognized linear epitope estimation representations using the structures Preferred and ML techniques. Hidden Markov Model (HMM), ANN, and SVM are the most often used machine learning algorithms. Physicochemical characteristics, as well as antigenicity scales, are commonly used in all of these methods.

Table1. *Comparison of ML methods in epitope estimation.*

Methods Name	Features	ML Technique applied
ABCPred	Hydrophilicity, accessibility, flexibility, turns, polarity and antigenicity	Feed Forward and recurrent Neural network
BCPred	Hydrophilicity, accessibility, AAP antigenicity scale, flexibility, turns and antigenicity	Subsequence kernel based SVM
BepiPred	Levitt secondary structure and Parker hydrophilicity scale.	HMM
Cheng et.al method	Hydrophilicity, AAP antigenicity scale, flexibility, turns, accessibility and antigenicity.	Support Vector Machine(SVM)
LEPS	Hydropathy, flexibility, turns, antigenicity, polarity, tripeptide & tetrapeptide antigenicity, dipeptide and accessibility.	Radial Basis Kernel based SVM
AAT-fs	Amino acid triplet (AAT) antigenicity scale	Radial Basis Kernel based SVM
BayesB	A dipeptide's relative position specific amino acid tendency.	SVM employing Bayes Feature Extraction
DLBEpitope	Using a hugevolume of linear B-cell epitope data and trial indication from the IEDB database, cooperative deep learning increased the performance of linear B-cell epitope prediction.	Ensemble deep learning/ feed forward deep neural network

3. EXPERIMENTAL METHODS

In order to conduct the predictive analysis on the approach of linear B-Cell and Non-B-Cell Epitopes is proposed. Figure 2 shows the workflow in proposed methodology.Data Acquisition & Loading, Data Pre-processing & Exploratory Data Analysis, Tokenisation & building sequences of words and Data segmentation to creation of training and test datasets are the main steps in implementing this work.

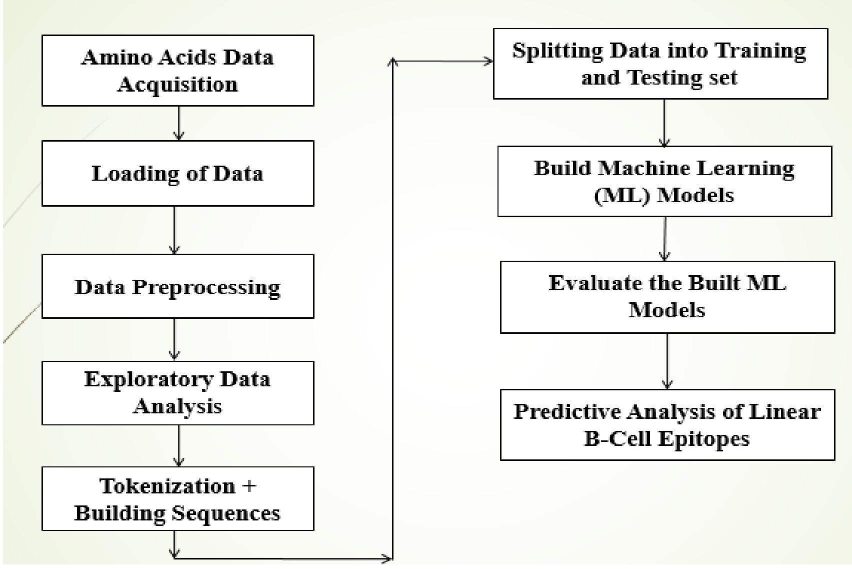
Amino Acids Data Acquisition
Loading of Data
Data Preprocessing
Exploratory Data Analysis
Tokenization + Building Sequences
Splitting Data into Training and Testing set
Build Machine Learning (ML) Models
Evaluate the Built ML Models
Predictive Analysis of Linear B-Cell Epitopes

Figure2. *Proposed Methodology*

3.1. Data Acquisition and Loading

The amino acids data are collected from open source data available [20]. To work with machine learning requires two things: data and models. When gathering data one can make sure that there are enough features (data elements that can aid in prediction) to properly train the learning model. In general, the more data one has, the better the outcome. The data is saved as a CSV file or another form of dataset that is supported. Loading the Data set includes the incoming data from the data source is merged with the current rows if the DataSet already contains rows. The Load function is useful in a variety of situations, all of which revolve around collecting data from a specified data source and adding it to the current data container (in this example, a DataSet). Table 2 shows the sample data acquired from open source [20].

Table2. *Sample data of Linear B – Cell epitopes*

Sl. No.	Amino Acids Sequence	Type
1.	YYVDKGHNYCGYPENLLIPK	Bcell Epitopes
2.	YYVPLGTQYTDAPSFSDIPN	Bcell Epitopes
3.	YYYADPEGPLPFPYFERQTI	Bcell Epitopes
4.	YYYDLSTIDPAEEIELQTIT	Bcell Epitopes
5.	YYTTTDSSSDSQTITNPAYD	Bcell Epitopes
6.	YYQQKPVALINNQFLPYPYY	Bcell Epitopes
7.	YYRENMHRYPNQVYYRPMDE	Bcell Epitopes
8.	YYRMMQTVRRMELKADQLYK	Bcell Epitopes
9.	YYTKNTNNNLTLVPAVVGKP	Bcell Epitopes
10.	YYTTTDSSSDSQTITNPAYD	Bcell Epitopes

In this work, data acquisition includes the loading of the B-cell and non-B-cell epitopes dataset, wherein dataset contains all the information of cells which are located in the amino acid that isNon-B cell and B cell

3.2. Data Pre-processing and Exploratory Data Analysis

In this phase, raw data is converted into a format that is useful and efficient. The processes involved in data pre-processing include data cleaning, data transformation, and data reduction.

- Cleaning the data: Many sections of the data (amino acids) may be irrelevant or missing. In order to handle this component, data cleansing is performed. It requires coping with data that is missing, noisy, and so on.
- Data transformation: The Steps taken to order the transformation of data (amino acid) into approximate forms by normalization, attribute selection, discretization.
- Data Reduction: Analysis gets more challenging when working with huge sets of amino acids. To get rid of this, this work uses a data compression approach. Its objective is to increase storage efficiency while simultaneously decreasing data storage and analysis costs.[21,22,23]

Exploratory Data Analysis: - Amino acid data analysis is the act of analyzing, cleansing, manipulating and modeling the amino acid data in order to identify usable information, draw conclusions and aid decision-making. As show in the below figure 3. The output of amino acid dataset after analyzing the B cell and Non B-cell, obtained an output of Non B – cell as 20,000 epitope and B – cell as 12,500 epitope.

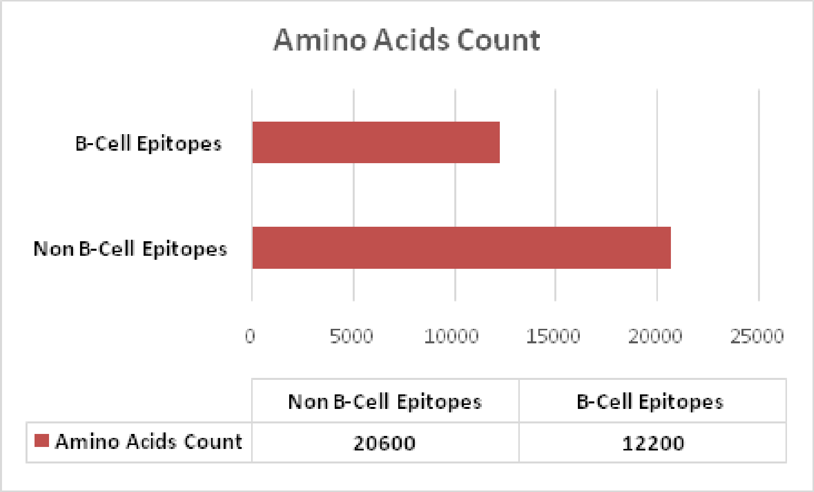

Figure3. *Exploratory data Analysis*

3.3. Tokenization and Building the Sequence of Words

Tokenization is the conversion of a string of characters into a string of tokens (strings with an assigned and thus identified meaning). Amino acid dataset represented in the form of A, B, C, D etc. are assigned to some specific numbers like 1, 2, 3, …….., so on. The figure 4 shows the frequency distribution of each amino acids.

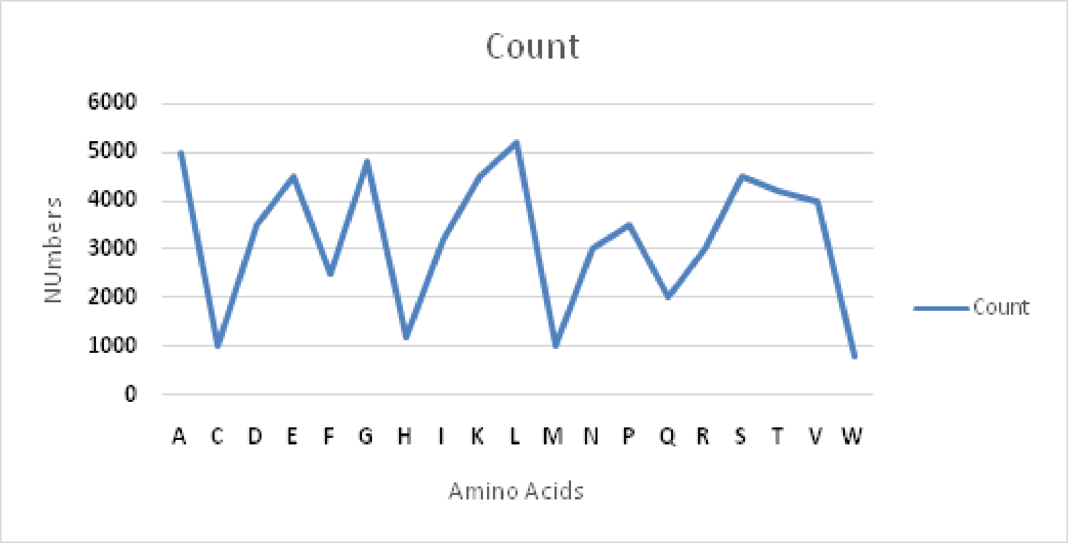

Figure4. *Frequency distribution of each amino acids*

3.4. Training and Test Datasets

Decision trees, random forests, logistic regression, and XGBoost are examples of machine learning modelsand Naive Bayes are selected for conducting the experiments with initial parameters set to default value.The pre-processed data is split into 80:20 training and test ratio and 70:30 training and test ratio to find the best training ratio to the given dataset. The model is trained in different split ratio and build models are stored in pickle file. So as to find the best fit of the given dataset, the hyper parameters are tuned for each machine learning build models.

4. RESULTS

This work is implemented in anaconda Jupiter notebook version 6.03 with python version 3.8.3. Initially tensor flow is configured and installed to get the utilities from python library. Then keras is installed for performing pre-processing operations. Sklearn package is used for loading the machine learning algorithms. After training the decision tree, NaiveBayes and logistic regression we found less in the accuracy of 63%. The performance is evaluated using metrics such as precision, recall and f1-score. To enhance the model accuracy, we tried tuning of hyper parameters such as random state, n_estimators and min samples split of random forest. The result obtained from random forest in the experiments is tabulated in table 3. This work explored XGBoost algorithm hyper parameters such as estimator rate, max_depth and gamma values. The obtained result from XGBoost is tabulated in table 4. The table 5 shows the overall comparative study results obtained. Graphically the results are compared and shown in figure 5.

Table3. *Results obtained from Random Forest*

Random_State	N_Estimators	Min_Samples_Split	Accuracy	Precision		Recall		Score(F1)	
				0	1	0	1	0	1
1	100	1	65.4	68.5	92.4	63.4	93.7	65.8	92.9
33	200	5	64.3	62.9	89.	41.9	94.7	49.8	91.4
56	300	10	63.0	61.0	88.5	43.2	94.0	50.5	91.2
85	400	10	65.8	64.5	89.5	46.9	94.4	54.5	91.7
101	500	15	64.7	65.7	88.2	40.4	95.9	50.4	91.6

Table4. *Results obtained from XGBoost*

Estimator Learning Rate	Max_depth	Gamma value	Accuracy	precision		recall		Score(F1)	
				0	1	0	1	0	1
200,0.075	3	0.0	65.5	65.15	91.4	59.4	93.8	62.5	92.9
300,0.1	4	0.1	64.2	63.11	89.6	45.9	95.0	54.2	91.9
400, 0.25	6	0.2	63.25	62.9	88.3	41.4	94.8	50	91.4
500,0.5	10	0.3	63.4	62.8	88.9	45.8	94.1	53	91.5
600,0.75	15	0.4	63.5	63.8	88.5	42.4	94.9	50.9	91.5

Table5. *Accuracy Obtained from build ML models*

Built ML Models	Accuracy (%)
Decision Tree	63.06
Bernoulli Naive Bayes	63.06
Logistic Regression	63.06
XGBoost	64.05
Random Forest (Proposed)	**65.78**

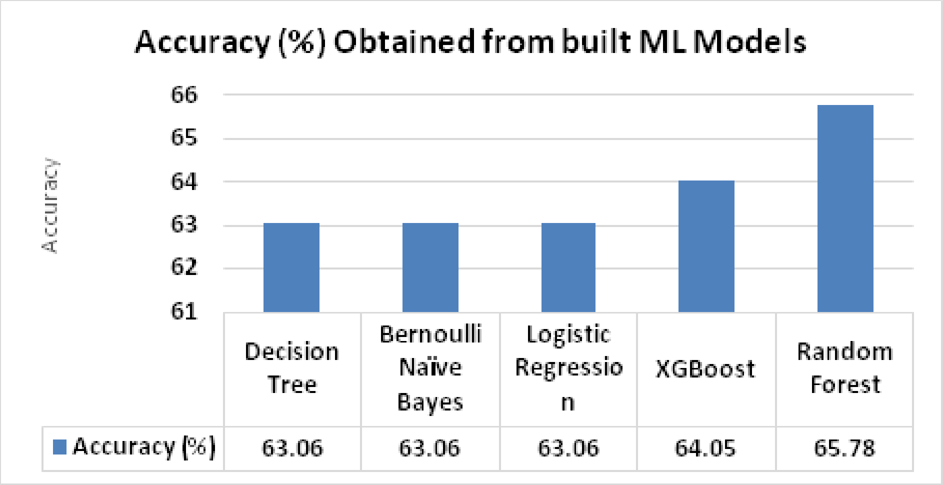

Figure5. *Accuracy obtained from ML Models*

5. Conclusion

In this predictive analysis study different machine learning models are explored and evaluated for linear and nonlinear B-cell epitopes. Feature collection stays significant in epitope estimation. Various kinds of hyper parameters are examined systematically to find the best predictive model for the given dataset. Performance accuracy differs depend on learning techniques used and features selected. Integrating the numerous features applied with some factors that aren't important for epitope prediction can be filtered out using all of these methods including Principal Component Analysis.

References

[1] https://www.cancercenter.com/community/blog/2017/05/whats-the-difference-b-cells-and-t-cells

[2] Hsin-Wei Wang and Tun-Wen Pai, “Machine Learning-Based Methods for Prediction of Linear B-Cell Epitopes”, July 2014, Methods in molecular biology (Clifton, N.J.) 1184:217-36, DOI: 10.1007/978-1-4939-1115-8_12

[3] Kosmas A. Galanis @el, “Linear B-cell epitope prediction: a performance review of currently available methods, doi: https:// doi.org/10.1101/833418

[4] EL-Manzalawy, Y., Honavar, V. “Recent advances in B-cell epitope prediction methods. Immunome Res 6, S2 (2010). https://doi.org/10.1186/1745-7580-6-S2-S2

[5] Chandra, S., Singh, T.R. Linear B cell epitope prediction for epitope vaccine design against meningococcal disease and their computational validations through physicochemical properties. Netw Model Anal Health Inform Bioinforma 1, 153–159 (2012). https://doi.org/10.1007/s13721-012-0019-1

[6] KavithaK V at el. “Computational Methods in Linear B-cell Epitope”, International Journal of Computer Applications (0975 – 8887), Volume 63– No.12, February 2013

[7] J.E.P.Larson, O.Lund and M.Neilsen, “Improved Method for predicting linear B-cell epitopes” Immunome Research. 2:2, 2006

[8] S. Saha and G. Raghava, “Prediction of continuous B- cell epitopes in an antigen using recurrent neural network.” Proteins, vol 65: pp 40-48, 2006.

[9] J.Sollner and B. Mayer, “Machine learning approaches for prediction of linear B-cell epitopes on proteins”. Journal of Molecular Recognition., vol 19, pp 200-208, 2006.

[10] J.Chen, H. Liu,J.Yang and K.Chou, “Prediction of linear B-cell epitopes using amino acid pair antigenicity scale”Amino Acids, vol 33, pp 423-428,2007.

[11] S. Saha and G. Raghava,”Bcipep: A database of B-cell epitopes”, BMC Genomics, Vol 6, pp 79, 2005.

[12] L.Wang, J.Liu, S.Zhu and Y.Y.Gao, “Prediction of Linear B-cell epitopes using AAT scale” Third International Conference on Bioinformatics and Biomedical Engineering, ICBBE, pp 1-4, 2009.

[13] Y.EL-Manzalawy, D.Dobbs and V.Honavar, “Predicting linear B-cell epitopes using string kernels”. J. Mol. Recognit., vol 21, pp 243-255, 2008.

[14] Y.EL-Manzalawy, D.Dobbs and V.Honavar, “Predicting flexible length linear Bcellepitopes”7th International Conference on Computational Systems Bioinformatics, pp 121-131, 2008.

[15] W.Zhang and Y.Niu,”Predicting flexible length linear B-cell epitopes using pair wise sequence similarity”, Third International Conference on Biomedical engineering and Informatics, 2010.

[16] L. JK. Wee, D.Simarmata,,Y. Kam, F P. Lisa and J. C. Tong “SVM-based prediction of linear B-cell epitopes using Bayes Feature Extraction” BMC Genomics, vol 11,Dec 2010.

[17] H.W.Wang,Y.C.Lin and H.T.Chang,” Prediction of B-cell linear epitopes with a combination of support vector machine classification and amino acid propensity identification” Journal of Biomedicine and Biotechnology, June 2011.

[18] Liu, T., Shi, K. & Li, W. Deep learning methods improve linear B-cell epitope prediction. BioData Mining 13, 1 (2020). https://doi.org/10.1186/s13040-020-00211-0

[19] Shastry K.A., Sanjay H.A. (2020) Machine Learning for Bioinformatics. In: Srinivasa K., Siddesh G., Manisekhar S. (eds) Statistical Modelling and Machine Learning Principles for Bioinformatics Techniques, Tools, and Applications. Algorithms for Intelligent Systems. Springer, Singapore. https://doi.org/10.1007/ 978-981-15-2445-5_.

[20] Hifzur Rahman Ansari, Harinder Singh, G. P. S. Raghava, LBtope : Prediction of Linear B-cell Epitopes: https://webs.iiitd.edu.in/raghava/lbtope

[21] Prakash, B. A., Ashoka, D. V., &Aradhya, V. M. (2012). Application of data mining techniques for software reuse process. Procedia Technology, 4, 384-389.

[22] Prakash, V. Ajay, D. V. Ashoka, and VN ManjunathAradya. "Application of data mining techniques for defect detection and classification." In Proceedings of the 3rd International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA) 2014, pp. 387-395. Springer, Cham, 2015.

[23] Praskash, B. A., Ashoka, D. V., Aradhya, V. M., &Naveena, C. (2016). Exploration of neural network models for defect detection and classification. International Journal of Convergence Computing, 2(3-4), 220-234.

Citation: *Mr. Adithya D A, Dr. Gowrishankar (2021). “Predictive Analysis of Linear B-Cell Epitopes in Immune Function using ML Algorithms”, International Journal of Research Studies in Computer Science and Engineering (IJRSCSE), 8(1), pp.23-30. DOI: http://dx.doi.org/10.20431/2349-4859.0801003*