Cardiovascular Disease Detection using Artificial Immune System and other Machine Learning Models

Ishan Gupta1, Ruchir Shangle1, Vishwas Latiyan1, and Umang Soni2
1 Netaji Subhas Institute of Technology, University of Delhi, New Delhi 110078, India
2 Netaji Subhas Institute of Technology, University of Delhi, New Delhi 110078, India

Email: ishan.gupta011@gmail.com

Abstract: Researchers and medical institutions face problems in detecting and diagnosing cardiovascular diseases in the early stages. Therefore, having a tool for detecting cardiovascular diseases in early stages will be helpful for the medical institutions to combat the disease. In this paper, we have presented a solution for detection of cardiovascular diseases by using clonal selection algorithm. Clonal selection is an Artificial Immune system (AIS) based algorithm which is often used for pattern recognition problems. Here, we propose modified clonal selection algorithm (CLONALG) which effectively detects cardiovascular diseases. With the proposed algorithm, we have achieved an average accuracy of 78%. Further, we compared the accuracy of CLONALG algorithm with different models of Machine Learning, viz. Random Forest Classifier (RFC), Decision Tree Classifier (DTR), Support Vector Machines (SVM), Logistic Regression (LR) and Artificial Neural Networks (MLP-ANN) for cardiovascular disease detection.

Index terms: Cardiovascular Disease Detection, Artificial Immune System, Clonal Selection Model, Artificial Intelligence, Machine Learning.

1. Introduction
Cardiovascular disease refers to the class of diseases that are encountered because of the inappropriate working of heart, blood vessels, or the related parts that are directly linked to the heart. CVD includes Coronary Artery Diseases (CAD) which includes Angina and Myocardial Infarction (or a heart attack). Cardiovascular disease is a prominent cause of morbidity and mortality worldwide. In 2016, nearly 17.9 million people lost their lives because of CVDs, which corresponds to 31% of all the deaths worldwide. In 2015, out of all 17 million premature deaths because of non-communicable diseases, 82% comprised low and middle-income countries, and CVDs caused 37% of it. People at high cardiovascular risk (because of risk factors such as diabetes, hypertension, hyperlipidaemia or already established disease) require an early diagnosis of symptoms and management using counselling and medicines, as appropriate. According to a recent estimate, 90% of all CVDs are preventable. Prevention involves eliminating risk factors that can be controlled naturally with minimal medication and cost through healthy eating and lifestyle, regular exercise, avoidance of tobacco products and smoking, and limiting alcohol intake. If carefully analysed, changing to a healthy lifestyle can prove to be effective against the prediction of CVD without excessive medicines which correspond to less treatment cost. People in third world countries rarely have the benefits of integrated health care programs for timely diagnosis and better treatment facilities compared to people in developed countries. People in most under-developed countries are diagnosed late in the course of the disease, which results in premature deaths, often in their most productive years. They affect the poorest of the poor the most. At the household level, because of
the vast health spending and high out-of-pocket expenditure for medical conditions, CVDs play a crucial role in the growth of poverty.

Thus, early detection of cardiovascular diseases plays an essential role in fighting against them. The biological immune system is highly convoluted and uses its adaptive nature and learning ability to fight foreign particles by maintaining a memory of similar or similar particles from past encounters. Taking the working and capacity of the immune system present in individuals into account, the Artificial Immune System was created, which was put in use to find solutions to various complex mathematical, technological problems [21]. We regard it hence as a computationally intelligent system that uses principles which form the basis of the working of immune system [15].

Originally, genetic and evolutionary computation techniques were the main idea for the building of AIS algorithm. Both the algorithms are modified version of evolutionary algorithms. The main difference between them is the mode of evolution of the population. Genetic algorithms [20] use crossover and mutation for the evolution whereas AIS is based on the asexual mode of reproduction in which each child is identical to its parent with a little mutation if needed. Mutation is required so as to ensure diversity in population and generating the best antibody cells.

The various algorithms in it are as follows: -

1.1 Negative Selection Algorithm
It works on self-no-self discrimination in the immune systems of mammals. In this, the self-reacting cells, which are T cells, are identified and deleted because they attack self-tissues. This process is known as adverse selection. We use this algorithm for problems related to pattern recognition and classification.

1.2 Dendritic Cell Algorithm
It is based on the principles and processes of the dendritic cells present in the immune system. The role of dendritic cells is to find damaged tissues and find the antigen responsible for this by traversing the tissues. Similarly, in DCA a population is created in which each agent works as a dendritic cell. These cells collect the data (antigen), which is used to put the antigen either under semi-mature cells or under mature cells. In binary terms, 0 is considered to be a normal antigen (semi-mature cell) and 1 as anomalous antigen (mature cell) [18].

1.3 Immune Network Algorithm
According to the hypothesis explained by Jerne et al. p [16], the immune system of organisms detects antigens with the help of an idiotypic network of interlinked B cells maintained by the immune system. In order to stabilize the net, the above-mentioned cells stimulate and suppress each other by specialized methods. This model was formerly named as the Artificial Immune Network (AIN), which was later updated and called Artificial Immune Network Algorithm (AINE) [12][17].

1.4 Clonal Selection Algorithm
It is based on the theory of the Clonal Selection of acquired immunity [11]. This theory explains the response of lymphocytes against the antigens attacking the body. In this, the antibodies that recognize the antigens undergo cloning, and their strength is increased through mutation. The above process helps in removing those with low affinities by replacing them with new antibodies. The whole process ensures that at the end, there is one antibody that is fully capable of binding to the antigen.

2. Literature Review
In the past couple of decades, they have developed many CVD classification models and algorithms, which mathematically incorporate different models to predict the risk of having/enacting a CVD, for instance - FRAMINGHAM, SCORE, and QRISK models. The Framingham Heart Study is acknowledged as one of the most extended and most crucial epidemiological studies in medical history. Initiated in 1948, the main objective of the Framingham Heart Study was to search for the risk factors and the effects of cardiovascular disease in a longitudinal population-based cohort. The enrolled 5209
residents of Framingham who were in the age group of 28 to 62 were examined. The examination consists of daily routine assessment, physical check-up, blood tests and 12-lead ECGs which were approved by the Investigational Review Board of Boston Medical Centre. The data was then interpreted and analytically studied to find general trends in cardiovascular disease patients. The study showed the ill effects of smoking and elevated blood pressure on the growth of CVDs in 1960s and 1970s [1].

Systematic Coronary Risk Evaluation (SCORE) is a study for CVD prediction which consist of high and low cardiovascular risk charts are based on factors which include patient’s gender and the age, blood pressure, total cholesterol, smoking status and other physical and medical features. SCORE also includes relative risk charts, identifiers and directions which are then studied analytically for the determination of CVDs. The SCORE dataset consist of 250,000 patient datasets from 12 cohort studies on the basis of which high and low risk charts have been developed and are excessively used in Europe. The biggest advantage of SCORE risk function is that it can be easily calibrated to each country’s national mortality statistics [2].

QRISK is a recent algorithm developed for CVD detection which works on the classical risk factor. These risk factors include age, systolic blood pressure, smoking status of the patient and ratio of total serum cholesterol to high-density lipoprotein cholesterol considering other risk factors such as body mass index (BMI), ethnicity, measures of poverty, family’s medical history and chronic kidney treatment. The results which were published in the BMJ showed the adroitness of QRISK algorithm over the classical Framingham model. Because of the unavailability of external validations, these models are inaccurate and biased when it comes down to practical applications.

Clonal selection algorithms [3] comes under a range of algorithms that are influenced by the clonal selection theory which is based on the Immune System of living organisms. The Clonal Selection Theory explains the enhancement of the response of B and T lymphocytes to antigens over time, which is known as affinity maturation. In Layman’s terms, it bases this algorithm upon natural clonal selection theory. We try to simulate the response of the immune system when an antigen invades the body. After the initiation of response, it generates a great diversity of antibodies to eliminate the antigens [4]. AIS CLONAL-G models are generally applied to the problems of pattern recognition [5] where they perform well. It is actively used in pattern recognition problems like the prediction of protein’s secondary structure [6] due to the versatility in choosing the required antibodies; so that the features of the problem can be effectively monitored by selecting the best antibodies for that respective feature. Nowadays this algorithm is modified for optimization of dynamic functions using a cluster based clonal selection approach [7]. In this study, we did some modifications in the traditionally used algorithm to study the efficiency of these class of algorithms in the classification domain [15].

3. CLONAL-G Algorithm
The proposed algorithm is based on how the actual human immune system responds in case when some infection/antigen enters the human body. The Immune system selects a particular lymphocyte that is similar to the attacking antigen and binds itself to the antigen to kill it. The algorithm combines the concept of clonal selection algorithm and k-nearest neighbour (KNN). The clonal selection algorithm produces memory cells for each class 1 & 0 where 1 represents that cardiovascular disease is present and 0 represents that cardiovascular disease is absent. These memory cells act like the lymphocytes, i.e., for each antigen we check that which memory cell is similar to the antigen and accordingly we group the antigen to that class. We use the idea of density and Euclidean distance to decide if the antigen is similar to the memory cell or not.
Figure 1. Clonal-G algorithm

Following is the detailed procedure for CLONAL-G algorithm:

- Antigens are selected from the dataset.
- The number of required antibodies is determined and is generated for the first generation.
- For one generation, affinity for all the antibodies is calculated for an antigen.
- Antibodies with top affinity are sent for cloning.
- The more the affinity of the antibody, the more the number of clones produced.
- The cloned antibodies are then forwarded for mutation.
- The rate of mutation for each cloned antibody is inversely proportional to its affinity.
- Affinity is calculated again for the mutated cloned antibodies, and top n antibodies are selected for that antigen.
- Antibodies with an affinity greater than those in memory cells for a particular antigen will replace those in memory cells.
- The above steps are repeated for every antigen.
- A set of updated memory cell cells was obtained.
- The above steps are repeated for a predetermined number of generations.
- For classification, the affinity of the antigen is calculated with the antibodies present in the memory cell, and the output is then generated depending upon which antibodies are closest to the antigen.
4. Description of different Machine Learning Models

4.1 Decision Tree Classifier
A decision tree classifier [22] works on the fundamental principle of creating a categorization model by building a decision tree. Every node in the tree defines a test on a feature; each branch descending from that node correlates to one of the potential values for that feature. For classification problems, it uses categorical variable decision trees.

Some important terms that are necessary for understanding decision trees are as follows:

- Root Node: It represents the original population which is further branched into two or more homogeneous subsets.
- Decision Node: When a sub-node further gets divided into various sub-nodes, it’s called a Decision node.
- Terminal Nodes: Nodes at the end of the tree that cannot be further divided are called Terminal nodes.
- Splitting: The action of the division of a node into its sub-nodes is called splitting.
- Pruning: The act of removal of sub-nodes of a decision node is called pruning.

4.2 Random Forest Classifier
Random Forest [23] is based on ensemble modelling. It is an estimator that takes various other estimators such as several decision tree classifiers on numerous subsets of the data and uses the mean value to improve the prophetic efficiency and control over-fitting. Random forest is an ensemble of a huge population of decision trees; each tree predicts something and the tree with the most voting wins, and its prediction is accepted. The flowchart for the algorithm is shown in fig. 2.

**Figure 2:** Flow diagram of random forest classifier
4.3 MLP-ANN Classifier
MLP, or multi-layer perceptron classifier, is a subclass of the feed-forward artificial neural network. It uses a mathematical model of a biological perceptron and uses it for processing of the data through forward and backward propagation (fig. 3). During the feed-forward propagation, the activation function calculates the weights for the neurons in the next layer of the network, and it goes on until the initial prediction is made. Then, backpropagation takes place. Here, the error margin of the predicted and true value of the output is calculated and propagated backwards for updating the weights on each neuron. It gets traced back up to the first, later, and at last, another round of feed-forward propagation happens. This gives us the desired output with a marginal error.

![Figure 3: Life cycle of MLP-ANN](image)

4.4 Support Vector Machines
Support vector machines make a subset of supervised learning algorithms that are very effective and are actively used in outlier’s detection, classification, and regression problems. SVM separates unique
attributes of the data on various hyperplanes and selects the most optimal hyperplane based on the margin. The hyperplane with the highest margin gets selected, which helps in classifying new data attributes.

4.5 Logistic Regression
Logistic regression [23] is a categorization algorithm that predicts the possibility of categorically dependent variables. The major assumption of logistic regression is that all the dependent variables should be binary in nature. In logistic regression as well, an activation function is used to approximate the value of the weights and calculate the loss function. After minimization of the loss function, we get the desired output.

5. Data Description
This dataset has entries from 70000 patients, and each entry has 12 different features associated with it. These features are used for training the model and further using it for predicting whether the patient has cardiovascular diseases or not. The details in these features are either factual information (OBJECTIVE) or results of medical examination (EXAMINATION) or have been provided by the patients (SUBJECTIVE).

The dependent variable in the dataset is the Cardio Column. If the value in this column is 1, it means that the patient has cardiovascular disease, and if the value is 0, it means that the patient doesn’t have cardiovascular disease. Table: 1 displays the characteristics of all the features.

| Feature                    | Variable Type       | Variable | Value Type                      |
|----------------------------|---------------------|----------|---------------------------------|
| Age                        | Objective Feature   | Age      | int (days)                      |
| Height                     | Objective Feature   | Height   | int (cm)                        |
| Weight                     | Objective Feature   | Weight   | float (kg)                      |
| Gender                     | Objective Feature   | Gender   | categorical code                |
| Systolic blood pressure    | Examination Feature | ap_hi    | int                             |
| Diastolic blood pressure   | Examination Feature | ap_lo    | int                             |
| Cholesterol                | Examination Feature | Cholesterol | 1: normal, 2: above normal, 3: well above normal |
| Glucose                    | Examination Feature | Gluc     | 1: normal, 2: above normal, 3: well above normal |
| Smoking                    | Subjective Feature  | Smoke    | binary                          |
| Alcohol intake             | Subjective Feature  | Alco     | binary                          |
| Physical activity          | Subjective Feature  | Active   | binary                          |

6. Results and Discussions
The study was performed on the cardiovascular disease dataset which predicted whether a person has CVD or not. In this study, various models were applied, examined and studied over the dataset containing 70000 complete examples. These models were compared with CLONALG for Artificial
Immune System. 80% of the available data was used as training data, and the remaining data was used for cross-validation.

The main focus of the study was to modify the clonal selection algorithm into a classification algorithm and compare its performance with other machine learning models in the prediction of cardiovascular disease. Scikit-learn was used for performance evaluation of AIS and other models.

The evaluation of the models was done on the following three metrics:

- **Accuracy**: It is the ratio of the total count of correct predictions to the total count of the predictions made by the model.
- **Matthews Correlation Coefficient (MCC)**: MCC is considered to be one of the best metrics for classification models. It incorporates all the cells of the confusion matrix in its formula.

\[
    MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP) \cdot (TP + FN) \cdot (TN + FP) \cdot (TN + FN)}}
\]

where,\( TP = \) count of true positives, \( TN = \) the count of true negatives, \( FP = \) count of false positives and \( FN = \) count of false negatives.

- **ROC and AUC curve**: (Here, ROC = characteristics of the receiving operator and AUC = Area under the curve) The function of ROC is to provide a probability curve whereas AUC gives us the degree of separability. Combined together, ROC and AUC represent the capability of our model to distinguish between various classes.

\[
    ROC\_AUC\_score = \int_0^1 TPR(FPR^{-1}(x))dx, \text{ where } TPR = \text{positive rate which is True and } FPR = \text{positive rate which is False}
\]

The CLONAL-G model has been trained for 20 generations and has 30 antibodies. The criteria used for the decision tree model was ‘Ginni’ with a maximum depth of 10 and a random state of unity. Random forest models were also trained on the same criteria, with a maximum depth of 15 and the number of estimators being 100. The number of epochs for MLP-ANN model was set to 300 with a random state of unity. The best results for SVM model and logistic regression were found for the default parameters of Scikit-learn library. The metrics values obtained for different models are shown in the table 2.

It is usually observed that the AIS CLONAL-G model works best for pattern recognition [10]. in this paper, it is evident from table 2 that the CLONAL-G model stands out from all the other models when we compare them on the basis of Accuracy and ROC-AUC score. When the value of Matthews’ correlation coefficient was considered, Random Forest performed best with a score of 0.496. Random forest models performed nearly as well as Clonal selection models, lagging behind in accuracy by just 0.052. But when we consider the complexity of the available dataset, a 0.052 or 5.2% makes a huge difference.

It can be observed from the table that the worst performance was shown by the SVM model with an accuracy of just 0.599, followed by the decision tree model. It is further observed that both these models cannot compete with other models when the MCC values are considered. Logistic regression can be ranked somewhere in between the above models, The value of accuracy metrics and ROC.AUC scores are nearly equal. The ROC-AUC score is generally less than the accuracy metric, but it is more in the case of Logistic regression. However, we cannot compare these two metrics as both have their own qualities and ranges (accuracy lies between 0 and 1, while the ROC-AUC score takes values between 0.5 and 1. The Matthews correlation coefficient (MCC), a reliable statistical tool that provides a steep score only if the model has predicted satisfactory results for all the four confusion matrix classes [15], marks the advantage of using Random forest over the CLONAL-G model.

Figure 5 shows the effect of number of generations on the loss function for the CLONAL-G algorithm. The loss function used in the CLONAL-G algorithm is the sum of Root Mean Square Error
(RSME) for all the training examples. The graph shows that generation 1 has the highest loss because for this generation the antibodies used were randomly generated. The loss function’s value decreases as the antibodies become more similar to the antigens due to mutation and cloning done in each generation. The loss value decreases rapidly during the early generations because of larger difference between the antibodies and the antigens. As the algorithm develops after each generation, the rate of change of loss value decreases and the loss value becomes nearly constant after 15th generation and finally becomes constant after 18th generation.

Table 2. Metrics Value

| Models             | Accuracy | MCC  | ROC_AUC |
|--------------------|----------|------|---------|
| AIS CLONAL-G       | 0.784    | 0.488| 0.760   |
| Decision Tree      | 0.631    | 0.263| 0.631   |
| Random Forest      | 0.732    | 0.496| 0.729   |
| MLP-ANN            | 0.642    | 0.315| 0.628   |
| SVM                | 0.599    | 0.186| 0.587   |
| Logistic Regression| 0.697    | 0.429| 0.707   |

Figure 4. Comparison of accuracy of different models
7. Conclusions
In this paper, we proposed modified clonal selection algorithm for cardiovascular disease detection. In this approach we combined the idea of KNN (k Nearest Neighbour) with clonal selection algorithm for prediction of cardiovascular diseases. We achieved a consistent accuracy of 78% for detection purposes. We also compared the performance of our algorithm with the other machine learning algorithms. On comparing the results, CLONALG algorithm came out to be best suited for predicting cardiovascular disease at an early stage. This shows that clonal selection algorithm can also be used as an accurate algorithm for classification problems along with pattern recognition problems. The immediate future direction would be to automate the parameter of the size of memory cell which could help in increasing the accuracy of the algorithm.

References
[1] Kannel, W.B. and McGee, D.L., 1979. Diabetes and cardiovascular disease: the Framingham study. *Jama*, 241(19), pp.2035-2038.
[2] Adamkiewicz, K., Płatek, A.E. and Szymański, F.M., 2018. Risk assessment according to the SCORE risk chart—from history, through present, to the future. *Kardiologia Polska (Polish Heart Journal)*, 76(1), pp.63-68.
[3] De Castro, L.N. and Von Zuben, F.J., 2000, July. The clonal selection algorithm with engineering applications. In *Proceedings of GECCO* (Vol. 2000, pp. 36-39).
[4] Timmis, J., Hone, A., Stibor, T., & Clark, E. (2008). Theoretical advances in artificial immune systems. *Theoretical Computer Science*, 403(1), 11-32.
[5] White, J.A. and Garrett, S.M., 2003, September. Improved pattern recognition with artificial clonal selection? In *International Conference on Artificial Immune Systems* (pp. 181-193). Springer, Berlin, Heidelberg.
[6] Yavuz, B.C., Yurtay, N. and Orkan, O., 2018. Prediction of protein secondary structure with clonal selection algorithm and multilayer perceptron. *IEEE Access*, 6, pp.45256-45261.
[7] Zhang, W., Zhang, W., Yen, G.G. and Jing, H., 2019. A cluster-based clonal selection algorithm for optimization in dynamic environment. *Swarm and Evolutionary Computation*, 50, p.100454.
[8] Forrest, S., Javornik, B., Smith, R.E. and Perelson, A.S., 1993. Using genetic algorithms to explore pattern recognition in the immune system. *Evolutionary computation, I(3)*, pp.191-211.
[9] Berek, C. and Ziegner, M., 1993. The maturation of the immune response. *Immunology today, I(48)*, pp.400-404.
[10] Silva, G.C. and Dasgupta, D., 2016. A survey of recent works in artificial immune systems. In *Handbook on Computational Intelligence: Volume 2: Evolutionary Computation, Hybrid Systems, and Applications* (pp. 547-586).
[11] Timmis, J., Neal, M. and Hunt, J., 2000. A survey of recent works in artificial immune systems. In *Handbook on Artificial Immune Systems* (pp. 547-586). Springer, Berlin, Heidelberg.
[12] Watkins, A., Timmis, J. and Boggess, L., 2004. Artificial immune recognition system (AIRS): An immune-inspired supervised learning algorithm. *Genetic Programming and Evolvable Machines*, 5(3), pp.291-317.
[13] Brownlee, J., 2005. Clonal selection theory & clonalg-the clonal selection classification algorithm (csca). *Swinburne University of Technology*, p.38.
[14] Hunt, J., Timmis, J., Cooke, D., Neal, M. and King, C., 1998. Artificial immune systems and their applications, chapter JISYS: development of an artificial immune system for real world applications.
[15] Dasgupta, D., 2006. Advances in artificial immune systems. *IEEE computational intelligence magazine, I(4)*, pp.40-49.
[16] Jerne, N.K., 1974. Towards a network theory of the immune system. *Ann. Immunol., I25*, pp.373-389.
[17] Hunt, J., Timmis, J., Cooke, E., Neal, M. and King, C., 1999. Jisys: the envelopment of an artificial immune system for real world applications. In Artificial Immune Systems and their Applications (pp. 157-186). Springer, Berlin, Heidelberg.

[18] Silva, G.C., Caminhas, W.M. and Palhares, R.M., 2017. Artificial immune systems applied to fault detection and isolation: A brief review of immune response-based approaches and a case study. Applied Soft Computing, 57, pp.118-131.

[19] Chicco, D. and Jurman, G., 2020. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC genomics, 21(1), p.6.

[20] Whitley, D., 1994. A genetic algorithm tutorial. Statistics and computing, 4(2), pp.65-85.

[21] Hofmeyr, S.A. and Forrest, S., 2000. Architecture for an artificial immune system. Evolutionary computation, 8(4), pp.443-473.

[22] Anyanwu, M.N. and Shiva, S.G., 2009. Comparative analysis of serial decision tree classification algorithms. International Journal of Computer Science and Security, 3(3), pp.230-240.

[23] Kleinbaum, D.G., Dietz, K., Gail, M., Klein, M. and Klein, M., 2002. Logistic regression. New York: Springer-Verlag