International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Novel Approach to Heart Disease Prediction using Machine learning

Krinal Thakkar[1], Dr Gayatri S Pandi[2]

[1] Post Graduate Scholar, Post Graduate Department LJ Institute of Engineering and Technology, LJ University, Ahmedabad
[2], Head of Department, Post Graduate Department, LJ Institute of Engineering and Technology, LJ University, Ahmedabad

ABSTRACT:

According to World Health Organization 32% of death is due to heart disease and 85% of death is due to heart attacks and stroke. This is mostly in the low and middle-income countries, In the USA someone is dying due to heart attacks every 40 seconds which has been a huge concern for heart disease prediction in it. Every year 805,000 people have heart attacks in the US. Many other diseases can also be responsible for heart disease and which can put people's health at high risk. In 2016 17.6 million people died due to cardiovascular disease which is 31% of global mortality. Each heart condition has different symptoms and other different warnings which can help a person to know about it. Most CVDs can stem from interference with normal heart function Conditions like obesity, blood pressure, and high cholesterol can contribute to such heart disease. All this can be prevented by proper diet, a good lifestyle, and proper sleep. Many people are given beta blockers for lowering their high blood pressure conditions. This model should answer complex answers with proper data which should be useful for the people, doctors and healthcare practitioners can make this cost-effective and should help in providing effective treatments. Machine learning is a very good way for the prediction of such diseases which will be very much helpful ineffective treatments. It provides the facility to improve the computer programs explicitly performed on it.

Keywords: Treatments, Machine learning models, Diagnosing heart disease, effective treatments

Introduction:

As per the modern world we are very much busy in this running race of earning money, and living a luxurious and comfortable life but due to this we are not taking care of ourselves and due to which our health is suffered a lot In this tension and pressure we are not having proper food habits and not a proper lifestyle, people nowadays do not have 8 hours of proper sleep and exercise due to which the body is affected a lot and we need to change our lifestyle otherwise we will suffer a lot. The quality of weather and food has also been changed nowadays due to which our heart and lungs do not get pure air and pure food and we start suffering from heart and lungs diseases

Machine learning is a type of Artificial Intelligence where that provides the facility to improve the computer programs and also the inputs and outputs given by the programs are based on the model and depending on the quality of the data processing made on the dataset Prediction means the relationship study between the response variable and predicted variable

There is a technique that should be performed on a response variable to predict the predictor variable and find different predictions of the same response. When a patient is having heart disease there are many factors for the same, factors can be Unhealthy food, hypertension, hypercholesterolemia, diabetes, tobacco, any major surgeries, etc. Heart attacks should be predicted months before with all these risk factors

Related Work:

The main aim of this study is to develop a prototype where we want to help the people and doctors by developing a healthcare prediction system that can bring the knowledge from hidden data in data storage, by providing such effective treatment it helps the people to reduce the costs of the treatment and we can visualize the data in tabular form and PDF forms

give a very good output in the health care, financial and stock market sectors. Machine learning model predictions allow business and private sectors to make highly accurate outcomes in the future and can also predict fraud, hacking and such criminal activities.

There are 14 attributes taken by the researchers for heart disease prediction and more than 300 records for it. They have used supervised learning techniques such as Naive Bayes, decision tree, random forest, KNN, etc. for predicting its accuracy of it. The dataset contains different types of attributes with their unit features. There are many types of datasets like the UCI dataset, and the Cleveland dataset This dataset contains 303 instances and 76 features, only 14 features are taken for testing the dataset and compared with different models like KNN, Random Forest, Naive Bayes, the important performance is of different algorithms

Classification of Machine learning techniques

The classification task is used for prediction and with the different comparisons of all algorithms used. Many machine learning models are used like Naive Bayes, KNN, Genetic Algorithm, and Random Forest with Linear model, - This model should provide a good performance and accuracy for the prediction of a dataset, we need to give an improvement in health care for better results, We can use Weka tools for dataset and more accuracy, preprocessing of data and graphs, there is a usage of different attributes taken from Weka tool and then the pre-processing is made on it. Only 16 attributes are taken from 83 attributes for testing.

Literature Study

The paper shows that the data samples are categorized into different steps In Step 1, the KNHANES-VI dataset was examined and data was selected. In step 2, statistical analysis was performed to identify features related to CHD risk. In step 3, predictors of CHD risk were selected using feature sensitivity-based feature selection in step 4, NN-based CHD risk predictors were trained using feature correlation analysis of features. In step 5, performance measurements were made to validate NN-based CHD risk predictions using feature correlation analysis.[2]

In this paper, we propose a novel method that aims at ending Signiant features by applying machine learning techniques resulting in improving the accuracy in the prediction of cardiovascular disease. [3] The prediction model is introduced with different combinations of features and several known class cation techniques. We produce an enhanced performance level with an accuracy level of 88.7% through the prediction model for heart disease with the hybrid random forest with a linear model (HRFM).[3]

In this study, I consider only 14 essential attributes. I applied four data mining classification techniques, K- nearest neighbour, Naive Bayes, decision tree and random forest. The data were pre-processed and then used in the model. K-nearest neighbour, Naïve Bayes, and random forest are the algorithms showing the best results in this model. I found the accuracy after implementing four algorithms to be highest in K-nearest neighbours (k = 7). We can further expand this research by incorporating other machine learning techniques such as time series, clustering and association rules, support vector machines, and genetic algorithms. [4] Considering the limitations of this study, there is a need to implement a more complex combination of models to get higher accuracy for the early prediction of heart disease.[4]

Cardiac surgery patients develop some form of AKI post-surgery 1–3and 1–5% develop severe kidney injury necessitating dialysis AKID.2,4–5 The mortality following AKI-D has been reported to be very high in the range of 50–80%.6,7 Even milder forms of AKI can have an impact on short term and long-term morbidity and mortality. AKI associated with cardiac surgery increases infectious risk and extends the length of stay in the intensive care unit thereby increasing the utilization of health care resources and independently predicting death.8 Recent advancements have led to less invasive surgical techniques and an off-pump

a system that can identify heart failure illnesses. For this purpose, our approach consists of developing a system that resolves the missing data problem in the preprocessing step by using the MICE model that we prove is the best algorithm to fill inexistent values.[5]

We used Boost, Ad boost, gradient boosting, extra trees, light gradient boosting Light gbm, SGDC, Nu SVM and the stacking algorithm in the classification step, the score accuracy of 95.83% was obtained by using the stacking algorithm. [5]

Cardiovascular diseases had been for a long time one of the essential medical problems. As indicated by the World Health Association, heart ailments are at the highest point of the ten leading reasons for death. Correct and early identification is a vital step in rehabilitation and treatment. To diagnose heart defects, it would be necessary to implement a system able to predict the existence of hair diseases. [5]

The perspective of said project is to make out a modernistic predictive model concerning the particulars of the eudaimonia for heart sickness based on predictive modelling. Numerous supervised machine learning techniques such as classification have been applied to existing therapeutic material. Different classification techniques will be implemented and compared upon standard performance metrics such as accuracy.[6]

A predictive system based on a hybrid intelligent machine learning technique was adopted for the treatment of heart sickness, which uses seven classifier algorithms namely LR, K-NN, DT, NB, ANN, SVM and RF with three different features selection methods, tested on Cleveland heart sickness dataset. In terms of accuracy, from all seven algorithms, the correctness result of logistic regression with the relief method provides an improved predictive outcome for heart sickness using a naïve Bayes classifier with a system for more enhanced predictions that was further developed by ensemble-based classifiers for more improvisation of system performance.[6]

Process Selection: Selection gives more copies to better individuals. But it does not always do so for better genes. This is because genes are always evaluated within the context of a larger individual. for example, consider the one max problem (that of counting ones). Suppose individual a competes with individual b When these two individuals compete, individual a will win. At the level of the gene, however, a decision error is made in the second position. That is, selection incorrectly prefers the schema *0** to *1**. The role of the population is to buffer against a finite number of such decision errors. Imagine the following selection scheme: pick two individuals randomly from the population, and keep two copies of the better one. This scheme is equivalent to a steady-state binary tournament selection. In a population of size, the proportion of the winning alleles will increase. For instance, in the previous example, the proportion of 1's will increase at gene positions 1 and 3, and the proportion of 0's will also increase at gene position 2. At gene position 4, the proportion will remain the same. This thought experiment suggests that an update rule increases a gene's proportion by simulating a small step in the action of a GA with a population of size The next section explores how the generation of individuals from a probability distribution mimics the effects of crossover.[7]

Process Crossover: The role of crossover in the GA is to combine bits and pieces from fit solutions. A repeated application of the most commonly used crossover operators eventually leads to a decorrelation of the population's genes. In this correlated state, the population is more compactly represented as a probability vector. Thus the generation of individuals from this vector can be seen as a shortcut to the eventual aim of crossover.[7]

Algorithms seem accurate and give evidence that the two are doing roughly the same thing and that they are somehow equivalent. Note however that while the SGA has a memory requirement of bits, the CGA requires only bits. The generation of individuals in the compact GA is equivalent to performing an infinite number of crossover rounds per generation in a simple GA. Thus the compact GA completely decorrelates the genes, while the simple GA still carries a little bit of correlation among the population's genes. Another difference is that the compact GA is incremental based while the simple GA is generationally based. One could get a better approximation of the two algorithms by doing a batch update of the probability vector once every competition is performed. This would more closely mimic the generational behaviour of the simple GA. We did not do that here because the difference is not significant. We are simply interested in showing that the two algorithms are approximately equivalent. This section modifies the compact GA that allows it to simulate higher selection pressures. We would like to simulate a tournament of size. The following mechanism produces such an effect.

1) Generate individuals from the probability vector and find out the best one. 2) Let the best individual compete with the other individuals, updating the probability vector along the way. The best individual wins all the competitions, thus the above procedure simulates something like a tournament of size Steps 2–4 of the CIA's pseudocode.

The aim of leading immune concepts and methods into GA is theoretically to utilize the locally characteristic information for seeking the ways and means of finding the optimal solution when dealing with difficult problems. To be exact, it utilizes the local information to intervene in the globally parallel process and restrain or avoid repetitive and useless work during the course, to overcome the blindness in an action of the crossover and mutation. During the actual operation, IGA refrains the degenerative phenomena arising from the evolutionary process, thus making the fitness of the population increase steadily. Because this course is very similar to that of an immune phenomenon in nature, the algorithm based on the above idea is named the IGA for simplicity and direct perception. To be more exact, the idea of immunity is mainly realized through two steps based on reasonably selecting vaccines, i.e., vaccination and an immune selection, of which the former is used for raising fitness and the latter for preventing deterioration.[8]

Proposed system

GA is a search heuristic approach, a process of natural selection where the fittest individuals are selected for the reduction to produce the offspring of the next generation, We need to make a proper sequential process of GA and start with the process of The natural selection which is the process of selecting the fitness of the features selected and then process it and create the next offspring, how better is this features will be as the offspring as how are parents as the children become, This is an iterative process at end generating the offspring and the fittest individuals should be found

The main phases are

1 Initial Population
2 Fitness function
3 Selection
4 Crossover
5 Mutation

1 Initial Population: The process where the individuals are called population, each individual is a set of variables and are called Genes, they join to form a chromosome
2 Fitness function: It displays how to fit the individuals and gives the scores to each individual, the probability that one should be selected is dependent on the fitness score
3 Selection: This is the phase where to select the fittest individual and then send them to the next generation. Two pairs should be selected and should be sent as per selected fitness scores and individuals with good fitness scores should have more chances to be selected.

4 Crossover: This is the most important phase in GA, where each pair of individuals are to be mated, a point is taken for random with the genes. Offspring are created by giving the genes of the fittest parents to each other until crossover is reached its point

5 Mutation: Here, new offspring are formed and some genes are given for the mutation with low probability, some of the offspring can be swapped if needed This is mandatory to maintain the diversity with the individuals and all coming offspring.

Encoding: It is the process of representing individual different genes.

The process can be performed by different methods

- Binary Encoding
- Hexadecimal Encoding
- Permutation Encoding
- Value Encoding
- Tree Encoding

Binary Encoding: Every Chromosome encodes a binary string. Every bit is a solution but not the best solution, the whole string can show or represent the number. It also differs from problem to problem. [20]

Hexadecimal Encoding: This encoding uses a string made up of hexadecimal numbers (0–9, A–F). [20]

Permutation Encoding: The chromosome should have a sequence of numbers, it should have some corrections to be made after genetic operations are completed. In this encoding, every chromosome should have real numbers. It is useful for order problems [19]

Chromosome A

1 5 3 2 6 4 7 9 8

Chromosome B

8 5 6 7 2 3 1 4 9

Value Encoding: It is the connected problem, form numbers, real numbers, or chars to some complicated objects. [19]

Tree Encoding: This is used for evolving programs and expressions, Each chromosome is a tree of some objects such as functions and commands of a programming language. LISP can be easily parsed as a tree, so the crossover and mutation can be done in a relatively easy way. [20]

Fitness Function: This can be assigned to evaluate the solutions. This quantifies the optimal solution. It depends on how the fitness value is close to the optimal solution. The fitness function determines how fit an individual is, it gives a fitness score to each individual. The probability that an individual will be selected for reproduction is based on its fitness score.

Selection: Selection is the process of choosing two parents from the population for crossing. After deciding on an encoding, the next step is to decide how to perform selection i.e., how to choose individuals in the population that will create offspring for the next generation and how many offspring each will create. The purpose of selection is to emphasize fitter individuals in the population in hopes that their offsprings have higher fitness.[20]

Types of Selection

1. **Roulette Wheel Selection:**

This is one of the traditional GA selection techniques. The principle of roulette selection is a linear search through a roulette wheel with the slots in the wheel weighted in proportion to the individual's fitness values, easier to implement but is noisy The rate of evolution depends on the variance of fitness's in the population. The roulette wheel will have a problem when the fitness values differ very much. If the best chromosome fitness is 90%, its circumference occupies 90% of a Roulette wheel, and then other chromosomes have too few chances to be selected. Consider a circular wheel. The wheel is divided into n pies, where n is the number of individuals in the population. Each individual gets a portion of the circle which is proportional to its fitness value. A fixed point is chosen on the wheel circumference as shown and the wheel is rotated. The region of the wheel which comes in front of the fixed point is chosen as the parent. For the second parent, the same process is repeated.

2. Random Selection

Randomly selects a parent from the population. Random selection is a little more disruptive, on average than roulette wheel selection. In this strategy, we randomly select parents from the existing population. There is no selection pressure toward fitter individuals and therefore this strategy is usually avoided.

3. Rank Selection

ranks the population and every chromosome receives fitness from the ranking. The worst has fitness 1 and the best has fitness N.It results in slow convergence but prevents too quick convergence. works with negative fitness values and is mostly used when the individuals in the population have very close fitness values (this happens usually at the end of the run). This leads to each individual having an almost equal share of the pie (like in the case of fitness proportionate selection) as shown in the following image and hence each individual no matter how to fit relative to each other has an approximately same probability of getting selected as a parent. This in turn leads to a loss in the selection pressure toward fitter individuals, causing the GA to make poor parent selections in such situations.

4. Tournament selection

In K-Way tournament selection, we select K individuals from the population at random and select the best out of these to become a parent. The same process is repeated for selecting the next parent. Tournament Selection is also extremely popular in literature as it can even work with negative fitness values.

5. Steady-state selection

This is not a particular method of selecting parents. The main idea of this selection is that a big part of chromosomes should survive to the next generation. In every generation, a few (good - with high fitness) chromosomes are selected for creating a new offspring. Then some (bad - with low fitness) chromosomes are removed and the new offspring is placed in their place. The rest of the population survives the new generation.

Crossover: Crossover is the process of taking two parent solutions and producing from them a child (offspring). A crossover operator is applied to the mating pool with the hope that it creates a better offspring. The reproduction operator selects at random a pair of two individual strings for the mating. A cross-site is selected at random along the string length. Finally, the position values are swapped between the two strings following the cross-site.

Crossover Operations

Although one-point crossover was inspired by biological processes, it has one major drawback in those certain combinations of schema cannot be combined in some situations. A multipoint crossover can be introduced to overcome this problem. As a result, the performance of generating offspring is greatly improved. An example of this operation is depicted where multiple crossover points are randomly selected[21]

Another approach is the uniform crossover. This generates offspring from the parents based on a randomly generated crossover mask. The operation is demonstrated. The resultant offspring contains a mixture of genes from each parent. The number of effective crossing points is not fixed, but will be averaged at L/2 (where L is the chromosome length) [21]

The preference of which crossover techniques to use is arguable. However, It concluded that a two-point crossover seemed to be an optimal number for multipoint crossover. This has since been contradicted by [loll as two-point crossover could perform poorly in a situation where the population has largely converged because of reduced crossover productivity. This low-crossover productivity problem can be resolved by the proposal of reduce-surrogate crossover[21] The different GA parameters are selection, mutation, and crossover. These are the most winning parts of the algorithm because they are contributing to the generation of each population. Selection phase-In the selection phase population individuals with superior fitness are selected, or else it is damaged. Crossover- In this method, a pair of individuals randomly participate in exchanging their parent’s genes with each other, awaiting an entirely new population to be generated. Mutation- It flips a number of the bits in an individual so that all bits are filled, and there is a low probability of predicting the altar[24]

Mutation: The Mutation function changes the chromosome in children in a preset probability (called a probability of mutation). In this application, the mutation function only mutates between phases within one intersection. The reason is to maintain the cycle time in each intersection. For example, one child has a chromosome of: [p11, p12, p13, p14, p21, p22, p23, p24,..., pn1, pn2, pn3, pn4] We then randomly select: a) an intersection u b) two phases v and w from this intersection and c) the variation (Var) within the range of (0, puv). The new duration of phases v and w are calculated as: pI UV = puv − V ar, puw = puw + V ar. The rest of the phase durations of this child remain the same.[26]

The mutation is an operator that maintains the genetic diversity from one population to the next population. The well-known mutation operators are displacement, simple inversion, and scramble mutation. The displacement mutation (DM) operator displaces a substring of a given individual solution within itself.

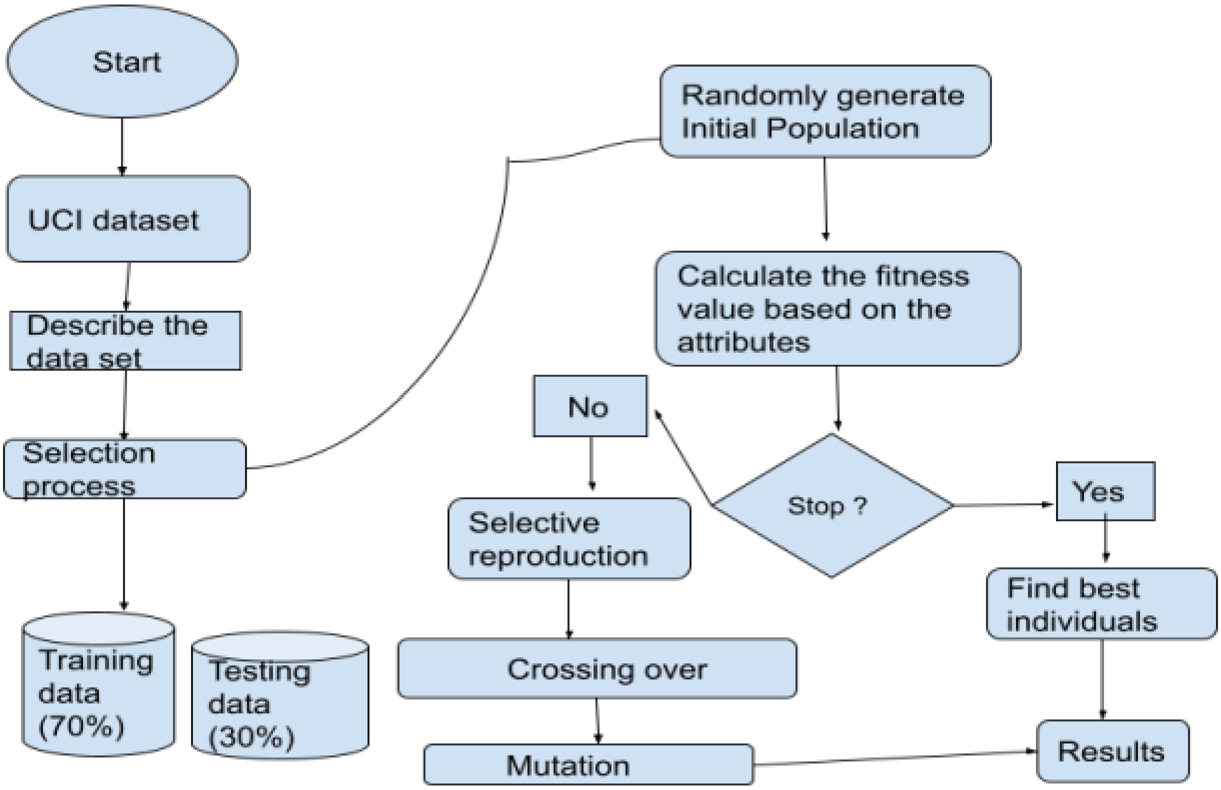

Figure:1 Flow of Genetic Algorithm

Conclusion:

This study will help us to know the strength and weaknesses of the other algorithms and models used and so we will be able to improve the prediction of heart disease by increasing the accuracy and performance of the model by getting the best dataset and the most efficient attributes in it which will provide the best accuracy, in model fitting the most important thing is feature selection and data pre-processing. The existing systems are focused on single attributes and single models, we can use more complex combinations of models and analyze and compare our results with other models, we should use more than 16 attributes and make a model based on the fitness functions given by genetic algorithm and send them for crossover process, we can also implement the number of training and testing data for more accurate results by the whole process of a genetic algorithm we can make the probability of the fitness function used by using Bayes theorem and make better results In this way it should be more accurate and can have good performance of the prediction methods used.

References:

[1] Malini Shukla Machine Learning Classification – 8 Algorithms for Data Science Aspirants Access from https://data-flair.training/blogs/machine-learning-classification-algorithms/ Access on 29/01/22

[2] Neural Network-Based Coronary Heart Disease Risk Prediction Using Feature Correlation Analysis Jae Kwon Kim and Sanggil Kang

[3] Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques Senthilkumar Mohan,
Chandrasegar Thirumalai, Gautam Srivastava

[4] Heart Disease Prediction using Machine Learning Techniques Devansh Shah, Samir Patel & Santosh Kumar Bharti Article number: 345 (2020)

[5] Machine learning-based identification of patients with a cardiovascular defect Nabaouia Louridi, Samira Douzi & Bouabid El Ouahidi

[6] Prediction of Heart Disease using Machine Learning Algorithms Garima Choudhary Dr Shailendra Narayan Singh Computer Science & Engineering

[7] The Compact Genetic Algorithm, Georges R. Harik, Fernando G. Lobo, and David E. Goldberg

[8] A Novel Genetic Algorithm Based on Immunity, Licheng Jiao, Senior Member, IEEE, and Lei Wang

[9]] Sonoo Jaiswal Javatpoint Access from - www.javatpoint.com, Access on 29/01/22

[10] Vincent Granville Towards Data Science Access from - https://towardsdatascience.com/ - Access on 29/01/22

[11] Sandeep Jain GeeksforGeeksAccess from https://www.geeksforgeeks.org/ Access on 29/01/22

[12] Heart Disease Prediction using Machine Learning Techniques by Vijeta Sharma,Shrinkhala Yadav,Manjari Gupta

[13] Virologist Dr Ijad Madisch Discover scientific knowledge and stay
connected to the world of science access from www.researchgate.net/ Access on 29/01/22
[14] Predicting Heart Disease at Early Stages using Machine Learning: A Survey by Rahul Kataria,
Polipireddy Srinivas
[15] Quincy Larson Learn to code — for free. Build projects. Earn certifications. Access from

www.freecodecamp.org, Access on 29/01/22
[16] Heart Disease Risk Level Prediction: Knitting Machine Learning Classifiers by Mrs Kelibone Eva Mamabolo Dr Moeketsi Mosia
[17] Michael Heinze DataSklr Access from www.datasklr.com Access on 29/01/22
[18]]Max Tegmark Becoming Human-Exploring Artificial Intelligence and what it makes to human Access from https://becominghuman.ai/ Access on 29/01/22
[19] Dr. Shu-Kun Lin Advancing Open Science for more than 25 years Access from /www.mdpi.com Access on 29/02/22
[20] Marek Obitko Access from https://www.obitko.com/tutorials/genetic-algorithms/encoding.php Access on 14/03/2022
[21] Genetic Algorithms: Concepts and Applications, K. F. Man, Member, IEEE, K. S. Tang, and S. Kwong,
Member, IEEE
[22] Genetic Algorithms in the Design and Optimization of Antenna Array Patterns, Francisco J. Ares-Pena, Senior Member, IEEE, Juan A. Rodriguez-Gonzalez, Emilio Villanueva-Lopez, and S. R. Rengarajan, Fellow, IEEE
[23] High Probability Mutation and Error Thresholds in Genetic Algorithms, Nicolae-Eugen Croitoru
[24] Intrusion Detection System using Fuzzy Genetic Algorithm, Yogita Danane, Thaksin Parvat
[25] Improvements in Genetic Algorithms, J. A. Vasconcelos, J. A. Ramírez, R. H. C. Takahashi, and R. R.
Saldanha
[26] Boosted Genetic Algorithm Using Machine Learning for Traffic Control Optimization, Tuo Mao,
Adriana-Simona Mihaita, Senior Member, IEEE, Fang Chen, and HaiL. Vu, Senior Member, IEEE
[27] A review on the genetic algorithm: past, present, and future, Sourabh Katoch, Sumit Singh Chauhan & Vijay Kumar
[28] An Ensemble Classifier Characterized by Genetic Algorithm with Decision Tree For the Prophecy of Heart Disease, K. Chandra Shekar, Priti Chandra and K. Venugopala Rao