# Feature Selection for neural network with genetic evolutionary algorithm

Abstract

## INTRODUCTION

Feature selection is a process of selecting a subset which can used for reducing the level of data by increasing the efficiency of the classification algorithm problem is an important task [2]. Fewer data, especially with particular fewer features, can often lead to more efficient calculations, simpler models, and better (e.g., more accurate) models. Data reduction is therefore necessary, even when the current computing power allows us to manage huge datasets because it can create advanced models [3]. A set of different integrated features should be obtained to keep the optimal combination to achieve optimal precision. In machine learning data mining and statistics. Evolutionary calculations (ECs) were created, based on some iterative evolution of the population of solutions that can solve optimization problems with some aspects of biological evolution. Evolutionary algorithms (EA) used the strategies of EC's. As EA follows EC's strategies thus like EC, EA also inspired by biological evolution. Several EAs used to select feature selection algorithm. Some of its strategies include principal component (PC), particle swarm (PS), ant colony (AC), and genetic algorithm (GA) optimizations techniques [3]. Recently, GA has been known as a very adaptive and efficient method of feature selection. The aim of this work is to apply two different machine learning methods (artificial neural networks and support vector machines), trained and structurally optimized by genetic algorithms, and to compare the results with regression-based methods (logistic regression).

## LITERATURE REVIEW

There are several published works for GA optimized classified models across different domains. For example, Barrios et al. [17] optimized NN topology and trained the network using GA and subsequently develops a classification technique for breast cancer classification. In another study, an SVM radial base function kernel parameter (width) was adapted to use GA as well as to select a subset of input features. In the second stage, the SVM classification technique was built and deployed to identify the machine conditions (defective or normal) [18]. Beiko and Charlebois [19] used GA to identify the best combinations of ANN topologies for DNA sequence classification. Karzynski et al. [20] used GA to optimize both architecture and ANN weight for classification in microarray classification. For instance, Cho et al. [22] used the ANN classification results as the GA fitness function in the cDNA microarray prediction. Bevilacqua et al. [23] and Lin et al. [24] applied error rates returned by ANN to determine the fitness of GAs in cancer classification.

## METHODS

### Study design

A cross-sectional survey (Multiple indicator cluster surveys, 2012) data (is one of the largest surveys conducted in Bangladesh) were used in this study, based on a sample of 51895 households (43474 rural, 8421 urban) interviewed with a response rate of 98.5%. The response variable is the school attrition. Attrition is identified by a woman, she attended school previously but was not in school in the current school year before data collection. Several factors are involved with the girl's school attrition. Predictor variables are included based on previous research [25]–[27].

### Data preparation

The data set was equally divided into two subsets: training and validation data. The main purpose of the validation data is to identify problems in the classification models such as overfitting effects or premature termination of the training process. This study used standardization of Z-scores to removes potential outliers of the data matrix. In the case of Z-score standardization, each attribute is subjected to standardization individually. In particular, the mean and standard deviations of a feature is computed. Then, the Z-score can be calculated by subtracting each element value as its mean and dividing it by the corresponding standard deviation. The formula to calculate z-value:

$$Z = \frac{X_i - \bar{X}}{s} \qquad (1)$$

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \qquad (2) \qquad\qquad s = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{s}} \qquad (3)$$

Where $X_i$ represents an element of each feature with "i" as an index. For a feature, the mean X and the standard deviation S are computed, respectively, N represents the total number of elements in a feature.

### A. Feature Selection Methods

Feature selection is a difficult issue not only because of the large search space but also because of feature interaction issues. Feature interactions occur frequently in many cases. Features can include many general way, sometimes, complex multiway interactions [5]. An individually relevant feature can do redundant work when used in parallelly with other features. Removing or selecting such features may miss the optimal feature subsets [2]. However, feature selection is used to select the key terms (features, e.g., words or phrases) in text mining [8], to create important graphical insides in image analysis [9], to find key genes from a large number of candidate genes in biological problems [6], and also to determine all important business indicators [7].

### B. Genetic Algorithm

The concept of GA (formerly known as genetic planning) was conceived by [10], as a method to explore a wide variety of optimization problems [11]. The design of the algorithm was inspired by observing the natural Darwinian evolutionary process and surviving the optimal principle [12]. Several of parameters need to be set of values while implementing GA, but the most important parameters are population size, mutation probability and crossover probability, and their interrelationships [13]. Genetic algorithm can be used for any optimization problem. It can also be used for training in logistic regression (LR), artificial neural network (ANN), and support vector machine (SVM) or determination of optimal performing structure. Since it is not using the error function of gradient to reach the best solution, it is also not sensitive to the local minima problem [14]. The evolutionary process of genetic algorithms is shown in Figure 1. First, some coded individuals are randomly generated and divided into initial populations. Second, each is given a fitness value by calculating fitness function, and individuals with high fitness are selected to participate in genetic operations while others are eliminated. Additionally, perform a cross-over, which is an exchange of features from the selected subset, then introduce mutations, which are applied randomly to the randomly selected features. Lastly, return to the second point. This continues until the exit criteria are found.

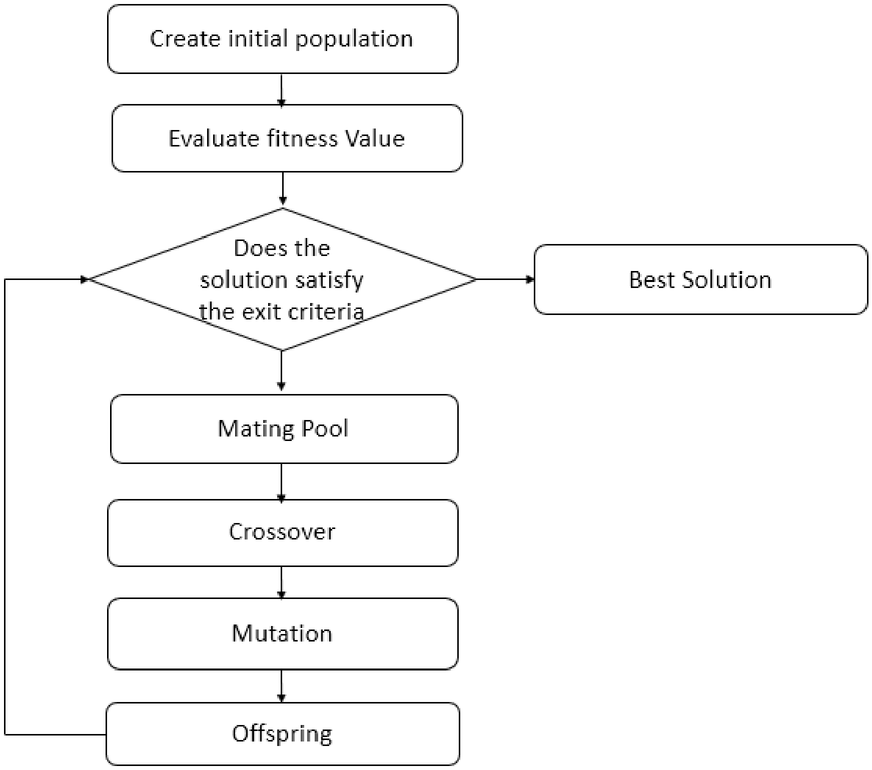


Figure. 1. Necessary iterative steps of the traditional genetic algorithm

**Initial Population:** The initial population and population size are the two major features in GA. Population initialization is known as the size of the population and usually selected by the criteria of the problem by initially generated random trial [15].

**Fitness computation:** Fitness computation are a key element of the classification model. We employ a 3-classification model to calculate the fitness value of the population. The fitness function of our model is defined as the number of properly labelled instances returned by the input samples of the models.

**Crossover:** This is the randomly pointed locus of an encoded bit string and the exact number of bits before and after the point position is broken down and exchanged between parental variables [16].

**Mutation:** It is the creation of an offspring from a single parent by inverting one or more randomly selected bits on the parent's variables. Mutations can be achieved with a small probability, for example, 0.001. The strings obtained from the crossover are mutated to avoid local minimums. Genetic materials that can be lost in the process of crossover and the distortion of genetic information are fixed through mutations. Mutation probability is responsible for determining how often the section will be the section of variables subjected to the mutation. Thus, the decision to mutate a section of the chromosome depends on the mutation probability. If the mutation is not applied, any part of the variables is generated immediately from the crossover without mutation. A 100% probability of a mutation means that the whole dataset will be changed, but a probability of 0% means that no part of the dataset will be distorted. Mutations prevent the GA from getting stuck at the local maximum [15].

Application of genetic algorithm are as follows:

1 **Initialization of population:** Generate an initial population that is a randomly generated bit string of binary values.
2 **Decoding:** Decode (bit string) to find out which input variable to select.
3 **Classification model:** In this stage, applied classification models.
4 **Fitness evaluation:** Take the prediction accuracy of each model as the fitness for GA.
5 **Stopping criterion:** The loop should continue or exit was determined here.
6 **Selection:** Select the model to cross over using randomly selected variables of the population.
7 **Crossover:** A linear combination of two selection was apply by a crossover operator.
8 **Mutation:** Attach a new gene with the same operator and create a random slot number of cross-over variables
9 **Replacement:** Replace old variables with the two-best combination which generate from mutation for next step.
10 **Loop:** Go to Step 2.

### Prediction Models

**Logistics Regression:** Logistic regression is the most common method for analysing binary response data. The outcome of the response variable Y is denoted by 1 ("success") or 0 ("failure"). The logistic model of any variable, Y, is defined as

$$\frac{1}{e^{-Y}+1} \qquad (4)$$

In logistic regression, whose equation is related to linear regression (Eq. 1), we try to find the values of the coefficients so that the fit of the data to the equation is maximized.

$$\text{Logit}\ (Y) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots\ldots\ldots \qquad (5)$$

where α and β’s are coefficient and X’s are variables.

Logistic regression can be generalized into categorical variables that fall into more than two possible categories.

**Support Vector Machine:** Support vector machine (SVM) is a classification tools that can deal with noisy pattens in a large dataset. The main goal of this supervisory approach is to find a function in a multidimensional space that can separate training data with known class labels. For example, if the data-points were plotted in N-dimensional space, a hyperplane could be detected that distinguishes the points belonging to one class from the points of another class.

**Artificial Neural Network:** ANNs got their name because they started as a simulation of how brain cells were believed to work. ANN provides an alternative method of interpreting and recognizing complex patterns in data sets. ANNs consist of one or more "layers" of autonomous computational units that receive input from other units and still transmit outputs to other units. Early ANNs were used for classification work, such as primary SVM or linear discriminant analysis. ANNs should be considered a form of converges that iterative itself to solve classification problems: it continues to be used in production for handwriting recognition, especially in US mail-where they enable automatic sorting.

**Fitness Evaluation:** The accuracy of the fitness values in GA were calculated as below:

$$\mathrm{Accuracy} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$$

Where TP is true positive, FP is false positive, TN is true negative, and FN is false negative. True Positive (TP) are those items that are properly selected. False positives (FP) are those items that have been selected incorrectly. True Negative (TN) are items that should not be selected but are selected. False Negatives (FN) are items that should be selected but not selected.

Performance criteria in this study was quantified by recall, precision and F1-score. The accuracy of how many selected items are acceptable and how many acceptable items are selected is called recall. A balanced blending of recall and precision is the F1-score. Performance metrics are also calculated in this study.

$$\mathrm{Recall(R)} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

$$\mathrm{F1\ score} = \frac{2\mathrm{PR}}{\mathrm{P} + \mathrm{R}}$$

## RESULTS AND DISCUSSION

In this study, the results were obtained based on the GA configuration, given in Table 1. Carefully selected fitness functions enable GA to reduce classification errors from classification models. As a proof, the best fitness and mean fitness should be close to the standard because GA has reached the condition of termination. Stall Generation is the number of generations produced by GA after the latest upgrade of fitness value. In this study, GA was used single point crossover, mutation threshold (0.55), maximum iteration (10), population number (40) and number of offspring in each iteration (40). T. Liu et al. worked previously with the population size of 20 and after 200 iterations all algorithms are stopped [30]. The parameters used by the GA for all models are shown in Table 1.

**Table 1: parameters used in GA**

| GA Parameter | Value |
|---|---|
| Crossover | Single point crossover |
| Mutation threshold | 0.55 |
| Max iteration | 10 |
| Number of population (Initial) | 40 |
| Number of offspring in each iteration | 40 |

Using the GA feature selector in ANN classification techniques, 6 features were reported with a mean of 0.67. Interestingly, GA has selected 7 features in both SVM and LR models. Ahmed et al. found the same feature with the use of similar parameters [31]. Since the parameters in the GA configuration table can still be fine-tuned for better results, the GA method has a higher level of controllability, with the mean feature value being higher in all models (Table 2).

**Table 2: Selected Feature**

| Feature | ANN | SVM | LR |
|---|---|---|---|
| Girls age (Year) | 1 | 1 | 1 |
| Marital status | 1 | 1 | 1 |
| Area | 1 | 0 | 0 |
| Divisions | 1 | 1 | 1 |
| Household wealth index | 1 | 0 | 0 |
| Religion | 0 | 1 | 1 |
| Household education | 0 | 1 | 1 |
| Mother alive | 1 | 1 | 1 |
| Father alive | 0 | 1 | 1 |
| Mean | 0.67 | 0.78 | 0.78 |

The training parameters of ANN are shown in Table 3. Currently there is no common method for selecting hidden layer nodes, learning rate and momentum rate. There is no standard for node selection when faced with different types of practical problems and different types of data. Experimental formulas are adopted to select these parameters. $M = \sqrt{}\ (n + l) + \alpha$ is the formula of calculating the hidden layer nodes, n and l represent input layer nodes and output layer nodes, respectively, α is an adjustable random variable between an 0-10. After much experimental effort for an optimal model, we found that a = 0, n = 49, l = 1 and the number of hidden layer nodes is 4 (6,9,9 and 4 hidden layer nodes, respectively). The learning rate generally values from 0.001; the momentum rate generally values from 0.6 to 0.8, here the research chooses 0.7. Regarding the learning rate chosen 0.01, the chosen 10 nodes for hidden layer and using 20% dataset for testing, have been reported in our previous study [32]. The selection of each parameter is a relative optimal solution obtained by multiple adjustments (Table 3).

**Table 3: parameters used in ANN**

| ANN Parameter | Value |
|---|---|
| Hidden layer | 4 [6,9,9,6] |
| Solver | adam |
| alpha | 1e-5 |
| activation | relu |
| output activation | softmax |
| Learning rate initialize | 0.001 |
| Maximum initialize | 500 |

It is clear that, without GA the prediction accuracy is lowest in all method. Without GA the accuracy was 76.00%, 75.25% and 74.92% in ANN, SVM and LR, respectively. On the other hand, With GA feature section method, the accuracy was 76.50%, 76.00% and 75.83% in ANN, SVM and LR, respectively (Table 4). Kim, used similar findings with us in their study. It is clear that, all method shows highest accuracy with GA [33].

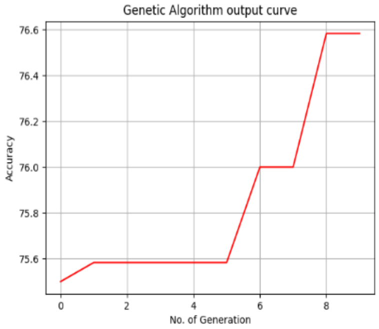
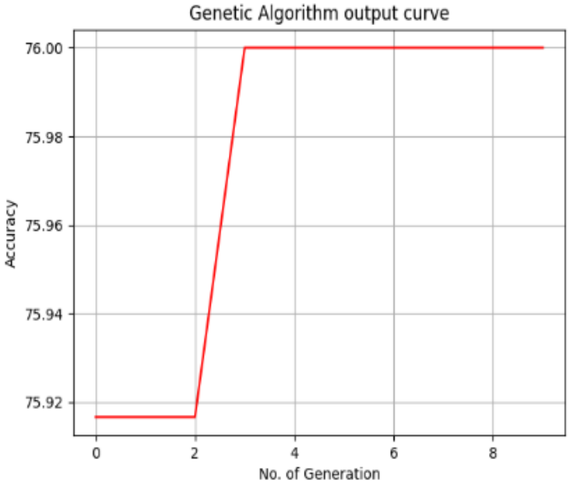
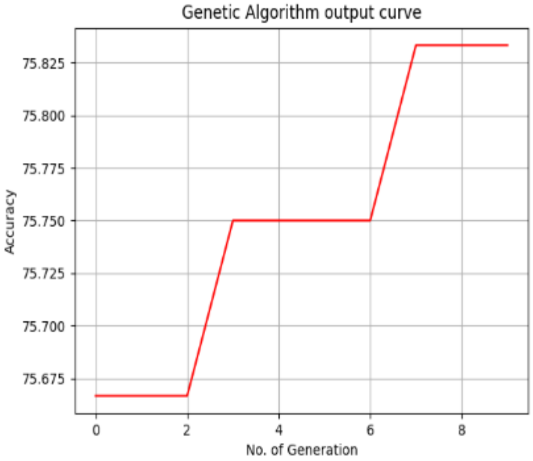
**Table 4: Classification accuracy**

| Classification Method | Without GA | With GA | Training time |
|---|---|---|---|
| ANN | 76.00% | 76.50% | 2.864 |
| SVM | 75.25% | 76.00% | 0.603 |
| LR | 74.92% | 75.83% | 0.043 |

We got the best recall of 95.39% in LR model which is better for this model than 50% and F1 score is highest on LR model and all method showed results above 50%. That means, all model fit best to our data.

**Table 5: Results of dropout classification using GA on different methods**

| Methods | Recall (R)<br>(%) | Precession (P)<br>(%) | F1-score<br>(%) |
|---|---|---|---|
| ANN | 92.35 | 77.65 | 84.36 |
| SVM | 93.67 | 76.59 | 84.27 |
| LR | 95.39 | 75.72 | 84.42 |

| | |
|---|---|
|  |  |
| | |
|  | |
| | |

Finding the best initial weight was a difficult task. It was not only about over-fitting from backpropagation in classification models, but also a tendency towards the minimal mean square error in training data without considering validation data in genetic algorithms. The advantages of using genetic algorithms compared to our previous research [32] should be based solely on the performance of the classification models of the testing datasets rather than the minimum square error in the modelling datasets. Another consideration for the minimal improvement of the genetic algorithm in this study was the number of subsets in a generation (i.e. the size of the population) and the length of the subset was small. This may be the reason why genetic algorithms cannot do extensive research to reach the best.

## CONCLUSION

In this study, we developed an individual classification method and GA model for classifying school attrition data. The results of the study show that the genetic algorithm got a good result for a small range of initial parameters. In most cases, the difference between the accuracy of the classification reported with and without GA is very small. Overall, the GA feature selectors created better classification accuracy without imply to this method. The main advantage of this approach is that it belongs to the field of controllable because GA can be secured for better results all the time by changing the fitness functions.

## References

[1] C. Emmanouilidis and C. Cox, "A Multi-Objective Genetic Algorithm Approach to Feature Selection in Neural and Fuzzy Modeling," vol. 3, no. 1, pp. 1–26, 2001.

[2] B. Xue, M. Zhang, W. N. Browne, and X. Yao, "A Survey on Evolutionary Computation Approaches to Feature Selection," *IEEE Trans. Evol. Comput.*, vol. 20, no. 4, pp. 606–626, Aug. 2016, doi: 10.1109/TEVC.2015.2504420.

[3] B. de la Iglesia, "Evolutionary computation for feature selection in classification problems," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 3, no. 6, pp. 381–407, Nov. 2013, doi: 10.1002/widm.1106.

[4] O. Babatunde, L. Armstrong, J. Leng, and D. Diepeveen, "A Genetic Algorithm-Based Feature Selection," *Int. J. Electron. Commun. Comput. Eng.*, vol. 5, no. 4, pp. 899–905, 2014.

[5] I. Guyon and A. M. De, "An Introduction to Variable and Feature Selection André Elisseeff," 2003.

[6] S. Ahmed, M. Zhang, and L. Peng, "Enhanced feature selection for biomarker discovery in LC-MS data using GP," in *2013 IEEE Congress on Evolutionary Computation, CEC 2013*, 2013, pp. 584–591, doi: 10.1109/CEC.2013.6557621.

[7] H. Liu and L. Yu, "Toward integrating feature selection algorithms for classification and clustering," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 4, pp. 491–502, Apr. 2005, doi: 10.1109/TKDE.2005.66.

[8] M. H. Aghdam, N. Ghasem-Aghaee, and M. E. Basiri, "Text feature selection using ant colony optimization," *Expert Syst. Appl.*, vol. 36, no. 3 PART 2, pp. 6843–6853, Apr. 2009, doi: 10.1016/j.eswa.2008.08.022.

[9] A. Ghosh, A. Datta, and S. Ghosh, "Self-adaptive differential evolution for feature selection in hyperspectral image data," *Appl. Soft Comput. J.*, vol. 13, no. 4, pp. 1969–1977, Apr. 2013, doi: 10.1016/j.asoc.2012.11.042.

[10] J. H. Holland, "Adaptation in Natural and Artificial Systems | The MIT Press," *The University of Michigan Press*, 1975. [Online]. Available: https://mitpress.mit.edu/books/adaptation-natural-and-artificial-systems. [Accessed: 31-

Jul-2020].

[11] H. Chiroma *et al.*, “Neural networks optimization through genetic algorithm searches: A review,” *Applied Mathematics and Information Sciences*, vol. 11, no. 6. pp. 1543–1564, 2017, doi: 10.18576/amis/110602.

[12] M. Hamdan, “A heterogeneous framework for the global parallelisation of genetic algorithms,” *Int. Arab J. Inf. Technol.*, vol. 5, no. 2, pp. 192–199, 2008.

[13] C. T. Capraro, I. Bradaric, G. T. Capraro, and T. K. Lue, “Using genetic algorithms for radar waveform selection,” in *2008 IEEE Radar Conference, RADAR 2008*, 2008, doi: 10.1109/RADAR.2008.4720947.

[14] “BP Neural Network Algorithm Optimized by Genetic Algorithm and Its Simulation,” *Int. J. Comput. Sci. Issues*, vol. 10, no. 1, pp. 516–519, 2013.

[15] S. N. Sivanandam and S. N. Deepa, *Introduction to genetic algorithms*. Springer Berlin Heidelberg, 2008.

[16] S. Haykin *et al.*, *Neural Networks and Learning Machines Third Edition*. 2009.

[17] D. Barrios, A. Carrascal, D. Manrique, and J. Ríos, “Cooperative binary-real coded genetic algorithms for generating and adapting artificial neural networks,” *Neural Comput. Appl.*, vol. 12, no. 2, pp. 49–60, Nov. 2003, doi: 10.1007/s00521-003-0364-1.

[18] B. Samanta, K. R. Al-Balushi, and S. A. Al-Araimi, “Bearing Fault Detection Using Artificial Neural Networks and Genetic Algorithm,” *EURASIP J. Appl. Signal Processing*, vol. 2004, no. 3, pp. 366–377, Mar. 2004, doi: 10.1155/S1110865704310085.

[19] R. G. Beiko and R. L. Charlebois, “GANN: Genetic algorithm neural networks for the detection of conserved combinations of features in DNA,” *BMC Bioinformatics*, vol. 6, no. 1, p. 36, Feb. 2005, doi: 10.1186/1471-2105-6-36.

[20] M. Karzynski, Á. Mateos, J. Herrero, and J. Dopazo, “Using a genetic algorithm and a perceptron for feature selection and supervised class learning in DNA microarray data,” *Artif. Intell. Rev.*, vol. 20, no. 1–2, pp. 39–51, Oct. 2003, doi: 10.1023/A:1026032530166.

[21] M. Taheri and A. Mohebbi, “Design of artificial neural networks using a genetic algorithm to predict collection efficiency in venturi scrubbers,” *J. Hazard. Mater.*, vol. 157, no. 1, pp. 122–129, Aug. 2008, doi: 10.1016/j.jhazmat.2007.12.107.

[22] H. S. Cho, T. S. Kim, J. W. Wee, S. M. Jeon, and C. H. Lee, “cDNA microarray data based classification of cancers using neural networks and genetic algorithms,” in *2003 Nanotechnology Conference and Trade Show - Nanotech 2003*, 2003, vol. 1, pp. 28–31.

[23] V. Bevilacqua, G. Mastronardi, F. Menolascina, A. Paradiso, and S. Tommasi, “Genetic Algorithms and Artificial Neural Networks in Microarray Data Analysis : a Distributed Approach,” *Eng. Lett.*, vol. 13, no. 3, pp. 335–343, 2006.

[24] T.-C. Lin, R.-S. Liu, Y.-T. Chao, and S.-Y. Chen, “Multiclass Microarray Data Classification Using GA/ANN Method,” Springer, Berlin, Heidelberg, 2006, pp. 1037–1041.

[25] M. N. Hasan, “Factors Associated with Attrition of Girls Students from School in Bangladesh,” *J. Sci. Res.*, vol. 12, no. 1, pp. 29–38, Jan. 2020, doi: 10.3329/jsr.v12i1.41579.

[26] M. N. Hasan, “A Comparison of Logistic Regression and Linear Discriminant Analysis in Predicting of Female Students Attrition from School in Bangladesh,” in *2019 4th International Conference on Electrical Information and Communication Technology, EICT 2019*, 2019, doi: 10.1109/EICT48899.2019.9068776.

[27] S. M. Shahidul and A. H. M. Z. Karim, “Factors contributing to school dropout among the girls: a review of literature,” *Eur. J. Res. Reflect. Educ. Sci.*, vol. 3, no. 2, pp. 25–36, 2015.

[28] F. A. Pujol, A. Jimeno-Morenilla, and M. J. Pujol, “Face Detection Based on Skin Color Segmentation Using Fuzzy Entropy,” doi: 10.3390/e19010026.

[29] T. Fawcett, “ROC Graphs: Notes and Practical Considerations for Data Mining Researchers,” *Pattern Recognit. Lett. Spec. issue ROC Anal. pattern Recognit.*, vol. 27, no. 8, pp. 882–891, 2006.

[30] T. Liu, H. Zhang, H. Zhang, and A. Zhou, “Information Fusion in Offspring Generation: A Case Study in Gene Expression Programming,” *IEEE Access*, vol. 8, pp. 74782–74792, 2020, doi: 10.1109/ACCESS.2020.2988587.

[31] F. Ahmad, N. A. Mat-Isa, Z. Hussain, R. Boudville, and M. K. Osman, “Genetic Algorithm - Artificial Neural Network (GA-ANN) hybrid intelligence for cancer diagnosis,” in *Proceedings - 2nd International Conference on Computational Intelligence, Communication Systems and Networks, CICSyN 2010*, 2010, pp. 78–83, doi: 10.1109/CICSyN.2010.46.

[32] Y. T. Chang, J. Lin, J. S. Shieh, and M. F. Abbod, “Optimization the initial weights of artificial neural networks via genetic algorithm applied to hip bone fracture prediction,” *Adv. Fuzzy Syst.*, 2012, doi: 10.1155/2012/951247.

[33] K. Kim, “Artificial neural networks with evolutionary instance selection for financial forecasting,” *Expert Syst. Appl.*, vol. 30, no. 3, pp. 519–526, Apr. 2006, doi: 10.1016/j.eswa.2005.10.007.