ISSN 2347 - 3983

Volume 8. No. 7, July 2020

International Journal of Emerging Trends in Engineering Research

Available Online at http://www.warse.org/IJETER/static/pdf/file/ijeter60872020.pdf

https://doi.org/10.30534/ijeter/2020/60872020

Diabetes and Heart Disease Prediction Using Machine Learning Algorithms

Syed Matheen Pasha [1], Shilpa Ankalaki [2]

[1]M.Tech Scholar Department of M. Tech CSE, Nitte Meenakshi Institute of Technology Bangalore, India,
syedmatheen602@gmail.com

[2] Assistant professor, Department of M. Tech CSE, Nitte Meenakshi Institute of Technology Bangalore, India,
shilpa.a@nmit.ac.in

ABSTRACT

Diabetes and Heart Disease are diseases with an ongoing illness that generates an augmentation and variation in the human body. Various troubles occur in case diabetes stays crude and unidentified. The dull perceiving handle occurred in going to of comprehension to a decisive focus and guiding expert. In any case, the ascent in Machine Learning approaches handles this essential issue. The reason for this is to predict the presence of diabetes as well as heart disease in patients with the most outrageous exactness. In this manner, ML counts to be explicit ANN, ELM, PCA, LASSO, Ensemble learning, and SVM was applied to recognize these diseases before it is orchestrated. The accuracy of the above mentioned ML algorithms is evaluated on Diabetes and Heart Disease Datasets.

Key words: Prediction of Diabetes, heart disease, Machine Learning, ANN, ELM, PCA, LASSO, Ensemble learning.
.

1. INTRODUCTION

Diabetes disease is an illness that occurs when your level of glucose is more. Blood glucose is the starting and major source which comes from the food which u have. Particularly a chemical made by the digestive gland makes dissimilarity sugar energy source from food get into your cells to be utilized for energy. Some of the times your corpse doesn't make enough or any upset or doesn't utilize offend well. Sometimes individuals call this disorder “a touch of sugar” or “borderline diabetes.” These conditions propose that somebody doesn't truly have diabetes or features a less genuine case, but each case of the disorder is serious. The mainly ordinary type of disorders is type 1, type 2, and developing diabetes. Sugar energy source at that point remains in your blood and doesn't arrive at your basic unit of a living thing.

Heart infection depicts a extend of conditions that influence your heart. Infections beneath the heart malady umbrella incorporate blood vessel maladies, such as coronary supply route infection; heartbeat issues (arrhythmias); and heart absconds you're born with (innate heart surrenders), among others. Machine Learning can play a fundamental part in foreseeing the presence/absence of Locomotors disarranges, Heart infections, diabetes, and more. Such data, in case anticipated well in development, can give vital bits of knowledge to specialists who can at that point adjust their determination and treatment per-patient premise. The contribution of this work is as follows:

- The proposed work employed the ML algorithms like ANN, ELM, SVM, LASSO, Ensemble learning for diagnosis of diabetic and Heart disease.
- Conducted the detailed experimental analysis of ANN and ELM by varying the Activation functions and learning rate
- Performance analysis of ANN, ELM, SVM, LASSO, Ensemble learning using Precision, Recall, and F1 score.

2. RELATED WORK

As explained by Nesreen Samer[1], The paper focuses on the neural network technique to analyze and demonstrate the diabetes prediction with better accuracy. As explained by Suyash Srivastava [2], Early determination can be made through a relatively reasonable strategy of computation. In this paper, Machine Learning algorithms are used to predict the early presence of diabetes. Classification is one of the foremost vital choice-making strategies for selecting data. In this, the most point of machine learning is to classify the information as diabetic or non-diabetic for progressing the classification exactness with the help of ELM. Pima Indian diabetes dataset taken from the UCI AI store is used. This dataset was browsed a greater dataset held by the National Organizing of Diabetes and Kidney Illnesses. The two-fold response variable takes the qualities '0' or '1,' where '1' infers a positive test for diabetes, and '0' could be a negative test for diabetes. The core of ELM is that the cultural boundaries of concealed center points, checking input loads and inclinations, are discretionarily allowed and require not to be tuned through the yield loads can be methodically chosen by the direct summed up turn around the activity. Diabetes is one of the preeminent regular contaminations around the globe where a cure is not found for it in any case. Yearly it incurred significant damage to a piece of money to think about individuals with diabetes. Subsequently, the preeminent basic issue is the desire to be especially exact and to use a strong procedure for that. Using LASSO the variable selection is done. As clarified by Suyash Srivastava1 [7], according to later examples heart disease has wound up the major ascertain for awkward passings. This survey habitat in giving the exactness of the model. The grouping computations, which have been breaking down join Naive Bayes, Random Forest, Extra Trees, and Logistic relapse which have been given chosen highlights utilizing least outright shrinkage and choice administrator (LASSO) and Ridge relapse. A group involves a lot of solely arranged classifiers, (for the model, neural systems, or variety trees) whose conjectures are joined even as ordering book occurrences. In this thesis, we weigh up these methodologies on 23 data sets using both neural frameworks and conclusion foliage as our characterization count. It comes about unmistakably show several conclusions. In any case, while Sacking is about ceaselessly more exact than a solitary classifier, it is now and again substantially less definite than Boosting. The paper [12] is a Machine Learning-Based Approach for the Identification of Insulin Resistance with Non-Invasive Parameters using Homa-IR. The paper [13] deals with the possibility of improving the performance speed of intelligent systems using the mathematical tool of artificial neural networks by introducing a control module of the genetic algorithm directly when performing the synthesis of solutions.

3. PROPOSED SYSTEM

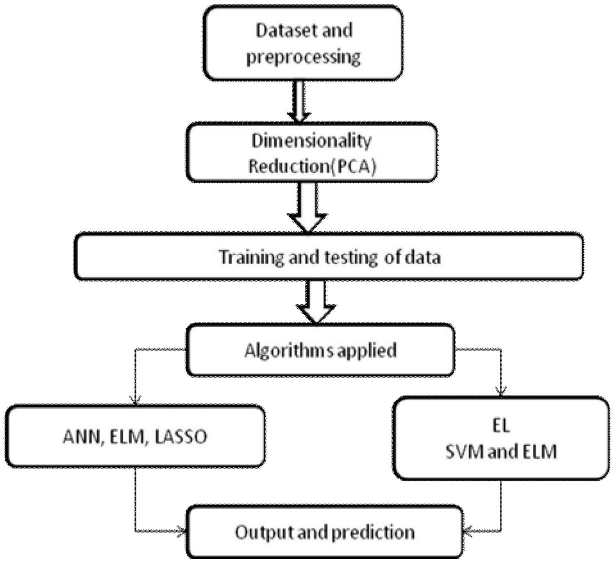
Dataset and preprocessing
Dimensionality Reduction(PCA)
Training and testing of data
Algorithms applied
ANN, ELM, LASSO
EL
SVM and ELM
Output and prediction

Figure 1: Proposed system

3.1: Dataset

The collection is initially from the nationalized organization, the goal of the dataset is to demonstratively look forward to whether an uncomplaining has sugar illness and heart disease, given assured indicative estimations remembered for the dataset. Independent variables include Age, Sex, Area of Residence, HbA1c, Height, Weight, BMI, duration of disease, another disease, Adequate Nutrition, Education of Mother, Standardized growth-rate in infancy, Standardized birth weight, Autoantibodies, Impaired glucose, metabolism, Insulin is taken, How Taken, Family History affected in Type 1 Diabetes, Family History, affected in Type 2 Diabetes, Hypoglycemia, the pancreatic disease affected in the child.

The heart dataset consists of 1026 individual data. There are 14 features in the dataset, which are age, sex, chest pain type, resting bp, serum cholesterol, fasting bp, resting ECG, max heart rate, angina, depression, ST segment, major vessels, thal, and diagnosis.

3.2: Preprocessing

Pre-processing is the modifications that are made to our data so that our information is fit to give as an input to the algorithm. Data that is cleaned is a method that cleans the raw information into a clean data form. The data which is collected from the unknown sources cannot be fed into the algorithms directly as it contains many impurities in it. Hence pre-processing is done to make the data clean.

With the help of Machine learning cleaning of the data is done. Initially, the data which is taken from the outside world i.e, from the unknown sources is been checked if there are any missing or unknown values present in it and they are replaced or removed so that the data looks fine. Then the null values are checked, as null values give no information which is the useless data, those are removed. Finally, as a resultant, we have data that is cleaned for taking it as an input to the machine learning models.

3.3: Dimensionality Reduction

The most immediate system for dimensionality decline, crucial part examination, plays out a straight mapping of the data toward a lesser gap so that the change of the data inside the low-dimensional portrayal is augmented. PCA is used for dimensionality reduction, where the dimensionality of the data is been reduced to a certain extent. This process is done to increase the accuracy of the model. After the application of PCA on some of the algorithms, the accuracy was increased. It does this by calculating the covariance of the data with the help of eigenvalues and vectors. Sort them in descending order and choose suitable vectors.

3.4: Training and testing of data

Training data and test data are two important concepts in machine learning. The dataset is divided into two-phase, one is training data and the other one is the testing set. Throughout the process, we divide the data into an 8:2 ratio i.e, 80% for the training phase, and the remaining 20% for the testing phase.

The dataset is divided as such for avoiding some complications such as overfitting and underfitting. The training set consists of the outcomes, on which the model learns to generalize the remaining data which can be used for further process. The other phase comes to the testing dataset, which is used to predict the output based on the dataset which is used for the training phase. Testing data is the main content with the help of which we can know the output of the model. But the testing dataset gives us the prediction only based on the training set since the ratio is large. If the data is not split up then there will be an issue such as generalization, underfitting, etc.

3.5 Algorithms

A. Artificial Neural Network(ANN)

Algorithm: ANN
Input: Dataset
Output: Predictions and accuracy
Parameters: Hidden neurons, learning rate, iterations, momentum.
Steps:

1. Get the input and convert it into arrays.
2. Every input is multiplied by its corresponding weight.
3. All the weights are then added up inside a unit.
4. If the sum is equated to zero then bias is added.
5. Pass the weighted sum to the activation function.
6. Desired output is generated using the activation function. The accuracy was boosted up by variation in the learning rate.

B. Multiple hidden layers Extreme Learning Machine(MELM)

Algorithm: MELM
Input: Dataset
Output: Predictions and accuracy
Parameters: Hidden neurons.
Steps:

1. At random assign weight between the participation and unseen node where the weight remains constant throughout.
2. compute the productivity matrix.
3. Produces the weight between the start and end node.
4. Finally uses the least square method to calculate the output weight of the network.

C. Least Absolute Shrinkage And Selection Operator(LASSO)

Algorithm: LASSO
Input: Dataset
Output: Predictions and accuracy
Steps:

1. Feature Selection: The best subset of features is picked utilizing a blend of the channel and covering incorporate assurance techniques.
2. Formulation: The LASSO-calculated backslide meaning of the issue is distinguished.
3. Initialization: The reproduced toughening show is instated using the formative strategy calculation.
4. Optimization Level: The boundaries (coefficients) of the LASSO model are streamlined utilizing a crossbreed formative method based reproduced fortifying system. We upgraded the boundaries of the proposed model.

5. Recognizing Solutions: We find the perfect game plan by looking at all arrangements.

D. Support Vector Machine (SVM)

Algorithm: SVM
Input: Dataset
Steps:

1. Input the diabetes information for classification.
2. Pre-process information to expel lost information.
3. It'll check the condition in case the information is classified at that point it'll show the predicted result.
4. If data isn't classified at that point it'll be classified into preparing information and testing information.
5. After this, it'll apply the SVM classifier for prediction.
6. Then it'll pre-process the information for removing the lost data. At that point, it'll show the anticipated result and find accuracy.

4. RESULTS AND ANALYSIS

4.1 Artificial Neural Network

The parameters used in ANN are Hidden neurons, learning rate, iterations, momentum. As there was an increase in the no of hidden neurons the accuracy boosted up with a learning rate changed to 0.09 range. Depending on the variations in the learning rate the model was good accurate for 0.06 value.

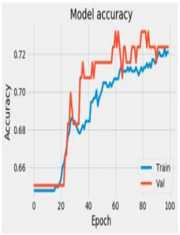

Figure 2: Model accuracy for ANN for Diabetic Dataset

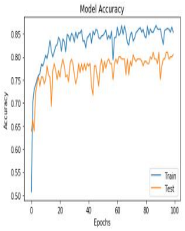

Figure 3: Model accuracy of ANN for Heart Disease Dataset

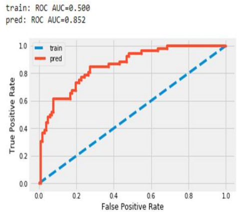

Figure 4: ROC curve obtained from ANN for Diabetic Disease Dataset

```
train: ROC AUC=0.500
pred: ROC AUC=0.107
```

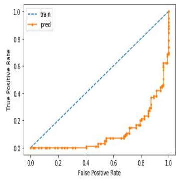

Figure 5: ROC curve obtained from ANN for Heart Disease Dataset

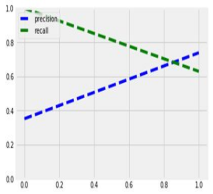

Figure 6: Precision-recall graph obtained from ANN for Diabetic Dataset

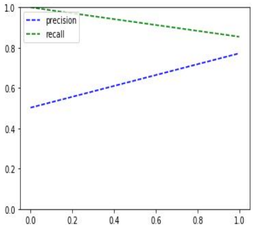

Figure 7: Precision-recall graph obtained from ANN for Heart Disease Dataset

Figure 2 and 3 depict the training and validation accuracy of ANN for various epochs for diabetic and heart disease classification. The model has achieved 75% and 85% accuracy for the diabetic dataset and heart disease classification. Figure. 4 and 5 show the ROC curve for ANN which gives the true positive rate plotted in function of the false positive rate (100-Specificity) for different cut-off points of a parameter. Fig. 2 depicts the training and validation accuracy of ANN for various epochs. The accuracy is dependent on the number of epochs we use. The model was 75% accurate based on the parameters. The ROC score was 50% for training and 85% for prediction. Figure 6 and 7 give the precision-recall graph which shows how well your model is in prediction. It gives the exactness of 78 for precision and 60 for recall.

4.2. Lasso

LASSO is used for variable selection and regularization techniques. After applying the LASSO algorithm on the diabetes dataset for prediction, the model was 79% accurate and for heart dataset, it was 80%.

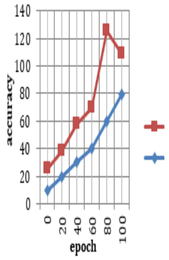

Figure 8: Model accuracy of LASSO for Diabetic Dataset

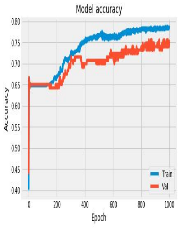

Figure 9: Model accuracy of LASSO for Heart Disease Dataset

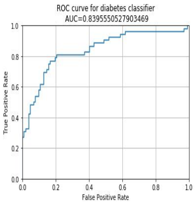

Figure 10: ROC curve obtained from LASSO for Diabetic Disease Dataset

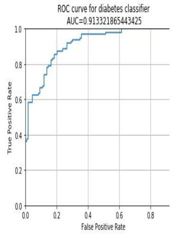

Figure 11: ROC curve obtained from LASSO for Heart Disease Dataset

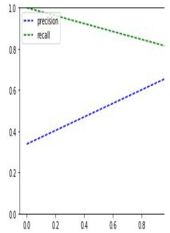

Figure 12: Precision-recall graph obtained from LASSO for Diabetic Dataset

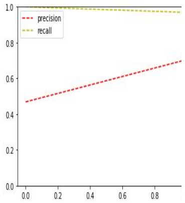

Figure 13: Precision-recall graph obtained from LASSO for Heart Disease Dataset

Figure 13 gives the precision-recall graph which shows how well your model is in prediction. It gives the exactness of 64 for precision and 79 for recall. Figure 12 gives the pr for heart data obtained from lasso for diabetes Dataset which depicts the performance of the model.

4.3. Support Vector Machine

SVM is supervised learning which is used for classification as well as regression analysis. On the diabetes dataset, the SVM model was 96% accurate and on the heart dataset, it was found to be 90% accurate. The corresponding graphs are plotted to analyze the result.

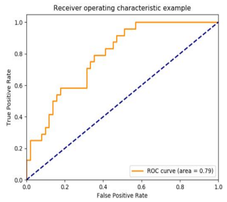

Figure 14: ROC curve obtained from SVM for Diabetic Disease Dataset

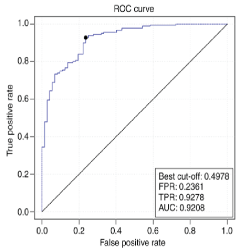

Figure 15: ROC curve obtained from SVM for Heart Disease Dataset

```
Average precision-recall score RF: 0.6971625554323726

Text(0.5, 1.0, '2-class Precision-Recall curve: AP=0.70')
```

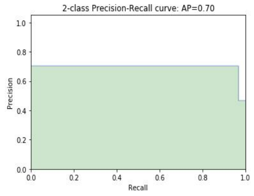

Figure 16: Precision-recall graph obtained from SVM for Diabetic Dataset

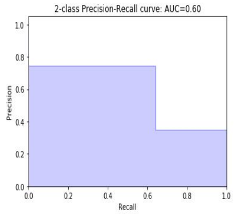

Figure 17: Precision-recall graph obtained from SVM for Heart Dataset

4.4 Decision tree

It is used for making decisions or decision analysis purposes. Fig 18 gives the ROC curve obtained from the Decision Tree for diabetic Disease Dataset. It gives the exactness of 72 for precision and 85 for recall as in fig 19. Figure 20 & 21 gives the pr obtained from DT for diabetes Dataset and heart data which depicts the performance of the model.

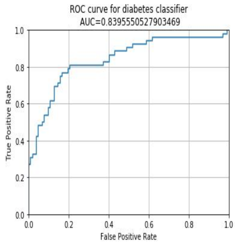

Figure 18. ROC curve obtained from Decision Tree for diabetic Disease Dataset

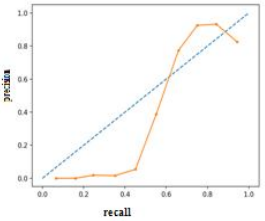

Figure 19. ROC curve obtained from Decision Tree for Heart Disease Dataset

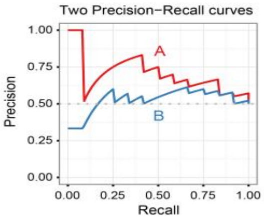

Figure 20: Precision-recall graph obtained from DT for diabetes Dataset

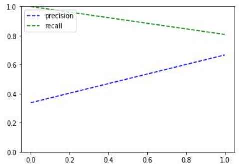

Figure 21: Precision-recall graph obtained from DT for Heart Dataset

4.5. Elm

On the diabetes dataset, the ELM with multiple hidden nodes the model was 72% accurate and on the heart dataset, it was found to be 70% accurate. The corresponding graphs are plotted to analyze the results. Figure. 22 and 23 depict the precision-recall curve and ROC curve of ELM for the diabetic dataset.

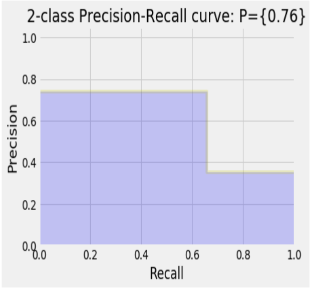

Figure 22: Precision-recall graph for Diabetes data

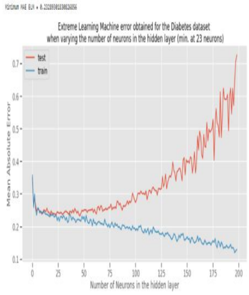

Figure 23: ROC curve for ELM

4.6. Ensemble learning

EL does average or voting to find the best accuracy of the model. On both the datasets the accuracy was similar that is less than or equal to 77%. The ROC curves and the precision-recall graph (figures 24,25,26) is plotted to analyze the result (figure 27).

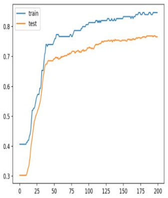

Figure 24: Model accuracy for EL

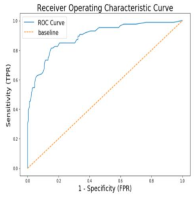

Figure 25: ROC curve for EL

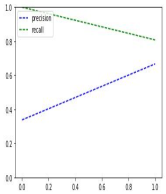

Figure 26: Precision recall graph for EL

Accuracy Graph:

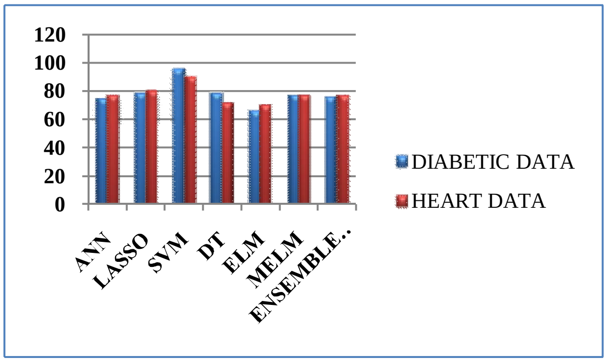

Figure 27: Accuracy of Machine Learning for diabetic and heart dataset

5. CONCLUSION

This is a comparative study of Machine learning algorithms. The classification was performed based on the no. of attributes or features that were present in the sample input dataset. The results obtained were efficiently represented using graphs. As the same input data was given for the comparison of the algorithms of results would be easy. The results obtained from the algorithms.

Lasso and Multiple extreme learning machines gave the highest accuracy compared to other algorithms.

ACKNOWLEDGEMENT

The authors express their sincere gratitude to Prof. N R Shetty, Advisor, and Dr. H C Nagraj, Principal, Nitte Meenakshi Institute Of Technology for giving constant support to carry out research at NMIT.

REFERENCES

1. Suyash Srivastava, Lokesh Sharma Vijeta Sharma, Dr. Ajai Kumar, And Dr. Hemant Darbari, **"Prediction Of Diabetes Using Artificial Neural Network Approach"**, *MANIPAL UNIVERSITY JAIPUR, ICOEVCI 2018, INDIA* https://doi.org/10.1007/978-981-13-1642-5_59
2. Dr. K. Anandha Kumar, Dr. A. Bharathi, Thiyagarajan C, **"Diabetes Mellitus Diagnosis based on Transductive Extreme Learning Machine"**, Bharathiar University, Department of Computer Science and Engineering, Department of Information Technology, Coimbatore. *International Journal of Computer Science and Information Security* (IJCSIS), Vol. 15, No. 6, June 2017.
3. Razieh Sheikhpour1*, Mehdi Agha Sarram, **"Diagnosis of Diabetes Using an Intelligent Approach Based on Bi-Level Dimensionality Reduction and Classification Algorithms",**

IRANIAN JOURNAL OF DIABETES AND OBESITY, VOLUME 6, NUMBER 2, SUMMER 2017

4. Shifei Ding, XinzhengXu, Rubie, "**Extreme learning machine and its applications.",** *Neural Comput&Applic, IJCSIS*, (2018) 25:549–556 https://doi.org/10.1007/s00521-013-1522-8
5. K. Hornik**, "Approximation capabilities of multilayer feedforward networks,"** *Neural Networks, Neurocomputing*, vol. 70, pp. 489-501, 2006. pp. 251—257, 2017.
6. M. Leshno, V. Y. Lin, A. Pinkus, and S. Schocken, **"Multilayer feedforward networks with a nonpolynomial activation function can approximate any function"** *Neural Networks*, vol. 6, pp. 861—867, 2016.
7. Gowda Karegowda Ashs1, Manjunath A.S, Jayaram M.S, **"Application Of Genetic Algorithm Optimized Neural Network Connection Weights For Medical Diagnosis Of Pima Indians Diabetes"**, *International Journal on Soft Computing (IJSC)*, Vol.2, No.2, May 2011. https://doi.org/10.5121/ijsc.2011.2202
8. Vijeta Sharma, Ajai Kumar, **"Outbreak Prediction model using Machine**", *International Journal of Intelligent Information and Management Science* (2018).
9. Dr. N. Ganesan, Dr.K. Venkatesh, Dr. M. A. Rama**, "Application of Neural Networks in Diagnosing Cancer Disease Using Demographic Data"**. *International Journal (2017).*
10. AvinashGolande, Pavan Kumar T, **"Heart Disease Prediction Using Effective Machine Learning Techniques**", *International Journal of Recent Technology and Engineering, 2018.*
11. Anisha.C.D and Dr. Arulanand.N**, "Early Prediction of Parkinson's Disease (PD) Using Ensemble Classifiers",** *International Conference on Innovative Trends in Information Technology (ICITIIT-2020).*
12. **Madam Chakradar, Alok Aggarwal**," **A Machine Learning Based Approach for the Identification of Insulin Resistance with Non-Invasive Parameters using Homa-IR**", *International Journal of Emerging Trends in Engineering Research*, vol. 8, no. 5, pp. May 2020. https://doi.org/10.30534/ijeter/2020/95852020
13. David Aregovich Petrosov, Roman Alexandrovich Vashchenko, Alexey Alexandrovich Stepovoi, Natalya Vladimirovna Petrosov," **Application of Artificial Neural Networks in Genetic Algorithm Control Problems**", *International Journal of Emerging Trends in Engineering Research*, vol. 8, no. 1, pp. 177-181, 2020. https://doi.org/10.30534/ijeter/2020/24812020