Assessment 1.2: Feature selection using stochastic search methods to Optimize Cardio Vascular disease predictions

Muhammad Ammar Shahid * [1]

Word Count: 2256

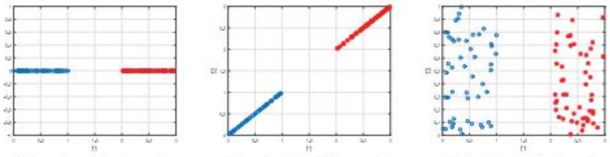
Abstract

In 2016, 31 percent of the deaths in the world were caused by Cardiovascular disease. An early detection of the disease can prevent the patient from death. Although there are machine learning models (ML) that can predicts the disease; but the scope of this report is to study the impact of stochastic search methods or Evolutionary Algorithms (EA) on existing solutions to increase the accuracy and minimize the features for better results. In this experiment three different EAs were used on a heart disease data-set on which many ML models are already trained. Still, decision tree is trained to see the impact of EAs to enhance accuracy and decrease the number of features. Results are remarkable, EAs achieved the both objectives i-e maximise the accuracy and minimise the features.

1. Problem Domain

In the era of big data, where the growth of high-dimensional data is increasing day by day in every field of life. Machine learning and deep learning play an important role in discovering knowledge automatically from data of various sorts. When machine learning algorithms are applied to high-dimensional data an issue known as the curse of dimensionality arises (Li et al., 2017).

A large number of irrelevant, redundant, and noisy features (Fig 1), may tend to over-fit the learning model, which can affect performance negatively. Likewise, high-dimensional data can crucially increase storage requirements and computational costs for training models and data analysis. To address the above-described issues, dimensionality reduction is considered one of the most powerful tools. It can be mainly categorized into two main components: feature extraction and feature selection(Li et al., 2017). This report only focuses on feature selection.



(a) relevant feature f_1 (b) redundant feature f_2 (c) irrelevant feature f_3

Figure 1. Feature f1 is a relevant feature that is able to discriminate two classes. However, given feature f1, feature f2 is redundant as f2 is strongly correlated with f1. Feature f3 is an irrelevant feature, as it cannot separate two classes at all. (taken from (Li et al., 2017))

Feature selection has been a research topic in methodology and practice for decades, is used by many fields, for instance, image recognition, text mining, bioinformatic data analysis and so on. Feature selection is a process, in which a subset of most relevant features are selected from the original set of features according to some feature selection criteria. Where redundant and irrelevant features are removed. A good feature selection technique can enhance accuracy, decrease learning time and reduce the complexity of the model (Cai et al., 2018). Feature selection techniques that are used in this report are discussed in the methodology 3.

2. Problem Instance

Cardiovascular disease (CVD), a class of disorders of the heart and blood vessels, is the most common cause of death in the world, representing 31 percent of all global deaths in 2016 (Roth et al., 2017). Early detection is mandatory to prevent CVD and can be done using ML techniques. Although a number of methods have already been used to predict the disease, a brief systematic review of ML approaches for CVD prediction is discussed by (Azmi et al., 2022). The review revealed the performance of different ML algorithms using different datasets in a tabular form and suggests that sample size has a very significant impact on the performance of the model. Another experiment on predicting CVD using

[1] M.Sc. Big Data Analytics, School of Computing and Digital Technology, Birmingham City University, UK. Correspondence to: <muhammad.shahid3@mail.bcu.ac.uk>.

Modern Optimisation Report, CMP 7213 - Modern Optimisation.

Table 1. Dataset columns and their description

Attribute	Description	Range
Age	Age of a person in years	29-79
Sex	Gender of a person (1-M,0-F)	0-1
CP	Chest pain type	1,2,3,4
Trestbps	Resting blood pressure in mm Hg	94-200
Chol	Serum Cholesterol in mg/dl	126-564
Fbs	Fasting Blood Sugar in mg/dl	0,1
Restecg	Resting Electrocardiographic results	0,1,2
Thalach	Maximum heart rate achived	71-202
Exang	Exercise induced Angina	0,1
OldPeak	ST depression induced by exercise relative to rest	1-3
Slope	Slope of the peak exercise ST segment	1,2,3
Ca	Number of major vessels colored by Fluoroscopy	0-3
Thal	3-Normal, 6-Fixed Defect, 7-Reversable Defect	3,6,7
Result	Class Attribute	0,1

a literature embedding model with the help of natural language processing achieved remarkable results(Moon et al., 2023).

This report uses the UCI heart disease data which is also available on Kaggle (SONY, 2020) and is cited in more than 50 publications. The UCI database contains 76 attributes but all publications refer to using a subset of 14 of them, attributes are described in Table 1. The same data is used by (Arghandabi & Shams, 2020) and built various ML models including logistic regression, Decision tree, SVM, gradient descent, and KNN. The experiment concludes that KNN outperforms all the other models with an accuracy of 85.7 percent.

The Objective of this report is to enhance the validation accuracy of Classifiers, using feature selection with the help of genetic algorithms. Minimizing the features without affecting the model accuracy will address the curse of high-dimensionality 1. Equation 1 describes the objective function or fitness function, which will return selected number of features and accuracy. A ML model would be trained on randomly selected features and fitness function will return model's accuracy multplied by number of selected features in negative (higher the number of features lower the fitness would be). Further details are discussed in 4

$$= -Feature_{min} * Accuracy_{max} \tag{1}$$

3. Candidate Optimization Methods

Genetic algorithm (GA) is mostly used for ML and problem optimization. Optimization is the technique that makes some existing solutions better. Optimization involves inputting values and obtaining outputs, which are used to create a set of all possible input and output pairs. This set represents the search space, which contains one or more points that can provide the optimal solution. The primary objective of optimization is to identify the point or set of points in the search space that yield the best results (Lambora et al., 2019). Working of GA is explained in Fig 2 Three different types of algorithms are used (Genetic Algorithm, Particle Swarm Optimization, and Multiobjective Optimization) to solve the problem 2.

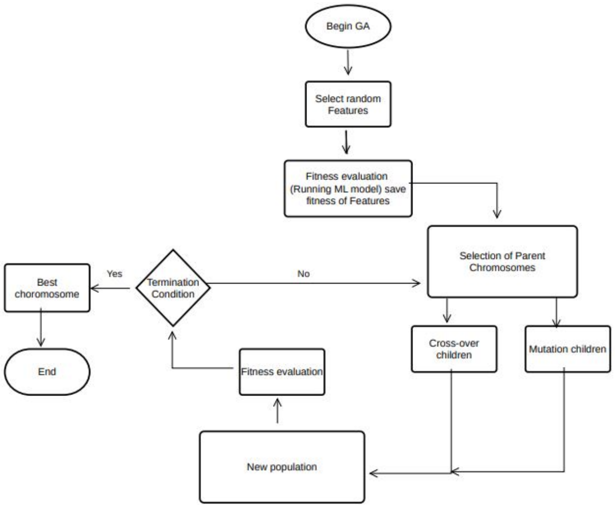


Figure 2. Working of GA explained (Modified from (Babatunde et al., 2014))

3.1. Genetic Algorithm

GA is a population-based optimization technique and algorithmic search heuristic method that is based on the natural evaluation process. GA produces a new population of chromosomes through an iterative procedure using genetic functionals such as cross-over and mutation as mentioned in (Babatunde et al., 2014). The same study used different algorithms to optimize the accuracy of concerned classifiers including GA, and GA outperformed the other Algorithms. This report will also use GA for the feature selection and to increase the accuracy of the CVD classifier using different crossovers, mutation, and population sizes and will compare the results with other algorithms used in this report.

3.2. Particle Swarm Optimization

Particle swarm optimization (PSO) is an EC technique proposed by (Kennedy & Eberhart, 1995). The PSO algorithm mimics the collective behaviour of animals such as birds and fish by representing the population of potential solutions as a swarm of particles in a search space. The algorithm begins by randomly initializing a group of particles, which then move around the search space in order to find the best solution. Each particle updates its position based on its own experience and the experiences of the particles around it, with the goal of collectively finding the optimal solution.

(Xue et al., 2014) discussed different approaches using PSO for feature selection, and a new method is also proposed. This report will implement the basic PSO to see its impact on the performance of the classifier.

3.3. Multiobjective Optimization Using NSGA-II

Since feature selection itself is a multiobjective problem (minimizing the features, maximizing the accuracy). (Abdollahzadeh & Gharehchopogh, 2022) compared different multiobjective optimization algorithms including NSGA-II for feature selection problems, and discussed them in brief. Furthermore, a hybrid algorithm based on two evolutionary algorithms was also used, which performed well on Feature selection problems.

This research will use NSGA-II on the dataset mentioned in 2 for feature selection and enhancing the accuracy of ML model. Results will be compared with GA and PSO.

4. Methodology

This section details the research design and techniques used in this study.

4.1. Data Pre-processing

Originally the data set that is used by authors (Arghandabi & Shams, 2020), have missing values, which were replaced by mean values of the respective columns. Another issue faced, while loading data was string values which can't be understand by the machine and were replaced by the numeric values using One-hotcoding method. Finally, in the target variable there were several outcomes (0-4). However actual results were only 2 i-e 0 for no heart disease and 1,2,3,4 for heart disease. so 2,3,4 were replaced by 1 as having different values for same outcome can effect the performance of the model.

4.2. ML model in Fitness function

Table 2. GA parameter selection

Population Size	20
Max Generations	30
Crossover (P)	0.8
Mutation (P)	0.1
Cross-Validation	Uniform
Convergence	20-runs

It was important to choose a ML model, which can return the fitness value of selected features in fitness function, for that purpose Decision-tree-classifier (DT) is used in this report. DT is a classification algorithm that can handle multiple classes in a dataset. When there are multiple classes with equal and highest probabilities, the classifier will predict the class with the lowest index out of all those classes. The authors (Arghandabi & Shams, 2020) also trained the DT and testing accuracy was 78.02 percent. This report focuses in the increase of testing accuracy and reduction of features as well to address the issue discussed in the 1. Hence fitness function will return the 1.

according to this equation accuracy is multiplied by the negative number of features. More the number of features more it will affect the accuracy negatively. Same fitness is used in all optimisation methods discussed in 3, to compare the effectiveness of each method.

5. Experiments

This report compared the results of different optimization methods that are previously discussed in 3 on a same fitness function that is discussed in 4. However selection of features would be random and could be different but 30 independent runs insures the best average results.

5.1. Dataset Description

Number of total observations in the data-set are 921, total number of male and female is 727 and 195 respectively. Average age of the patients is 54. 4 types of chest pain are observed in the data-set.

5.2. Parameter Selection

Scope of this report is to compare three different optimization methods and not to study the every method in detail. So parameter selection is same as default selection in R packages and kept remain same in all runs, as this report does not discuss the effect of change in parameter selection. However, seed values changed in every run to perform 30 independent runs.

In table 2 selected parameters are listed. With these parameters, 30 independent run with different seed values were performed. Seed value is set to get same random numbers for feature selection in future to get the same results. In

Table 3. PSO parameter selection

Swarm Size	20
Max Iter	20
Inertia Weight	0.7
Cognitive value	0.5
Social value	0.5
Bounds	0-1

Table 4. NSGA-II parameter selection

Population Size	20
Max Generations	30
Crossover (P)	0.8
Mutation (P)	0.1
Selection Operator	Non-dominated sort
Convergence	20-runs

each independent run, 30 generations were formed for 20 iterations. Population size shows the number of individual solutions.

Selected parameters for PSO are listed in table 3. Swarm represents the number of solutions per iteration. Inertia is responsible for how fast swarm moves towards the solution. Cognitive value is the individual's influence and social value represents other particle's influence. Bounds are set to 0 (feature not selected) for lower limit and 1 (feature selected) for upper limit.

Same parameters were selected for NSGA-II as GA, instead of cross-validation, NSGA uses parameter selection for that purpose non-dominated Pareto fronts were selected. see table 4

5.3. Results

Results achieved with EAs were remarkable and outperformed the results that the authors (Arghandabi & Shams, 2020) achieved. EM not only increased the accuracy but also reduced the features and address the issue discussed in 1. Fig 3 shows the comparison between three methods. It clearly shows that all the three methods achieved higher fitness value over the generations. In comparison with GA and PSO, NSGA-II outperformed both of them in achieving the higher fitness.

Fitness value should not be confused with accuracy, as mentioned in 4 fitness function not only returns the accuracy but also the number of features. So best fitness means, lower the number of features and higher the accuracy. GA and PSO achieved the fitness of 18.5 and 19.8 respectively, it can be seen in fig 3, that PSO achieved the result better than GA. Both achieved the max accuracy of 79.56 but PSO achieved on less number of features.

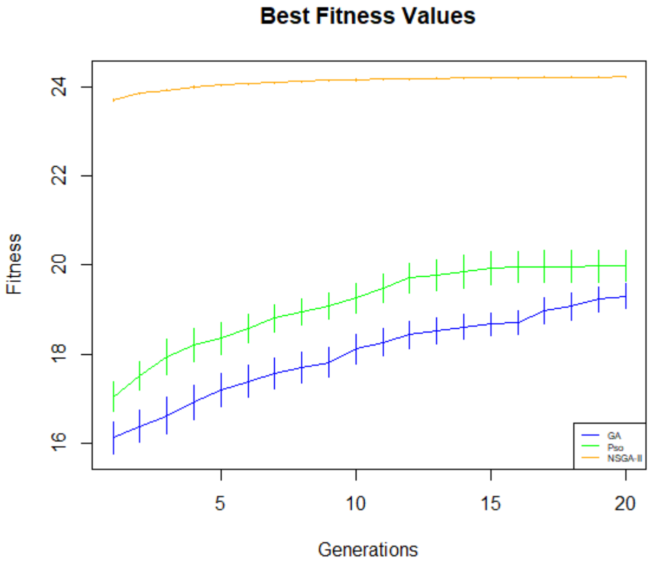


Figure 3. Comparison of GA, PSO and NSGA-II on fitness

Table 5. selected feature and accuracy for PSO and GA

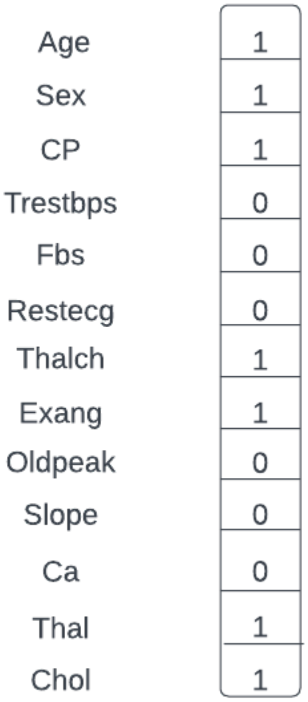
Condition	1	2	3
Accuracy	0.7956	0.7821	0.7778
Feature	0.60	0.50	0.60

Table 5 describes the selected features percentage and accuracy achieved with them. Maximum accuracy achieved with both GA and PSO is 79.56 percent which is higher than the accuracy achieved by (Arghandabi & Shams, 2020).

NSGA-II not only outperforms the (Arghandabi & Shams, 2020) results, but it also outperforms the other two EAs with the fitness of 24.22 and accuracy of 84.61 percent. Which shows that multi-objective optimisation (in this case 2 objectives i-e maximise the accuracy and minimise the features) can be performed better by NSGA-II which is a Multi-objective optimizer. Results are listed in table 6, and best solution in terms of selected features, which achieved highest accuracy can be seen in Fig 4. Where 1 indicates selected and 0 indicates un-selected features.

Table 6. selected feature and accuracy for NSGA-II

Condition	1	2	3
Accuracy	0.8461	0.8293	0.8043
Feature	0.53	0.65	0.70



Age	1
Sex	1
CP	1
Trestbps	0
Fbs	0
Restecg	0
Thalch	1
Exang	1
Oldpeak	0
Slope	0
Ca	0
Thal	1
Chol	1

Figure 4. Best solution (individual)

5.4. Discussion

The authors of (Barraza et al., 2017) discussed that NSGA-II performed better than PSO in their study, this report also achieved better results with NSGA-II, which clearly shows that NSGA-II is better than PSO and GA. The authors of (Abdollahzadeh & Gharehchopogh, 2022) compared the results of multi-objective optimizers (MO) and their study shows that NSGA-II achieved better result. However, This report only used one MO, and further experiments can be conducted to see the impact of different MO on the UCI heart disease data-set.

This report accurately classifies the heart patients with the accuracy of 84.61 percent, while using the DT in the fitness function. DT was randomly chosen to perform training and testing. Highest accuracy achieved by (Arghandabi & Shams, 2020) with DT was 78.02 percent with all the features. However, this experiment also reduced the number of features and are explained in the above 5.3. By reducing the features EAs also address the issue discussed in 1. Which can severely increase the computational costs and storage requirements. This report used a small data-set with 13 features and 921 observations, just to study the impact of EAs to solve the issue.

This report did achieve the better results and also address the issue of high dimensional data. It's impact can be surely seen on the data with much higher features. Training time, computational cost and storage needs can be decreased by using EAs. In addition to that models can achieve higher accuracy by removal of unnecessary features.

6. Conclusion

Machine lerning models can predict the disease in early stages, but don't have any method to remove the irrelevant features to enhance the accuracy. This study shows that EAs can do this for ML models.

Which can make the existing solutions more better and save training time, computational cost and storage needs. 5.3 shows that Within EAs NSGA-II is more suitable for optimising more than one objective (in this report minimize the number of features and maximize the accuracy). Predicting CVDs accurately with the less number of number features will also minimize the efforts to collect the more type of data. Just to clarify only the relevant feature would be collected in the future. Which can make the job even more easier.

References

Abdollahzadeh, B. and Gharehchopogh, F. S. A multi-objective optimization algorithm for feature selection problems. *Engineering with Computers*, 38(Suppl 3): 1845–1863, 2022.

Arghandabi, H. and Shams, P. A comparative study of machine learning algorithms for the prediction of heart disease. *Int J Res Appl Sci Eng Technol. https://doi.org/10.22214/ijraset*, 2020.

Azmi, J., Arif, M., Nafis, M. T., Alam, M. A., Tanweer, S., and Wang, G. A systematic review on machine learning approaches for cardiovascular disease prediction using medical big data. *Medical Engineering Physics*, 105:103825, 2022. ISSN 1350-4533. doi: https://doi.org/10.1016/j.medengphy.2022.103825. URL `https://www.sciencedirect.com/science/article/pii/S1350453322000741`.

Babatunde, O. H., Armstrong, L., Leng, J., and Diepeveen, D. A genetic algorithm-based feature selection. 2014.

Barraza, M., Bojórquez, E., Fernández-González, E., and Reyes-Salazar, A. Multi-objective optimization of structural steel buildings under earthquake loads using nsga-ii and pso. *KSCE Journal of Civil Engineering*, 21:488–500, 2017.

Cai, J., Luo, J., Wang, S., and Yang, S. Feature selection in machine learning: A new perspective. *Neurocomputing*, 300:70–79, 2018. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2017.11.077. URL https://www.sciencedirect.com/science/article/pii/S0925231218302911.

Kennedy, J. and Eberhart, R. Particle swarm optimization. In *Proceedings of ICNN'95-international conference on neural networks*, volume 4, pp. 1942–1948. IEEE, 1995.

Lambora, A., Gupta, K., and Chopra, K. Genetic algorithm- a literature review. In *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, pp. 380–384, 2019. doi: 10.1109/COMITCon.2019.8862255.

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., and Liu, H. Feature selection: A data perspective. *ACM Comput. Surv.*, 50(6), dec 2017. ISSN 0360-0300. doi: 10.1145/3136625. URL https://doi-org.bcu.idm.oclc.org/10.1145/3136625.

Moon, J., Posada-Quintero, H. F., and Chon, K. H. A literature embedding model for cardiovascular disease prediction using risk factors, symptoms, and genotype information. *Expert Systems with Applications*, 213:118930, 2023. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2022.118930. URL https://www.sciencedirect.com/science/article/pii/S0957417422019480.

Roth, G. A., Johnson, C., Abajobir, A., Abd-Allah, F., Abera, S. F., Abyu, G., Ahmed, M., Aksut, B., Alam, T., Alam, K., Alla, F., Alvis-Guzman, N., Amrock, S., Ansari, H., Ärnlöv, J., Asayesh, H., Atey, T. M., Avila-Burgos, L., Awasthi, A., Banerjee, A., Barac, A., Bärnighausen, T., Barregard, L., Bedi, N., Belay Ketema, E., Bennett, D., Berhe, G., Bhutta, Z., Bitew, S., Carapetis, J., Carrero, J. J., Malta, D. C., Castañeda-Orjuela, C. A., Castillo-Rivas, J., Catalá-López, F., Choi, J.-Y., Christensen, H., Cirillo, M., Cooper, L., Criqui, M., Cundiff, D., Damasceno, A., Dandona, L., Dandona, R., Davletov, K., Dharmaratne, S., Dorairaj, P., Dubey, M., Ehrenkranz, R., El Sayed Zaki, M., Faraon, E. J. A., Esteghamati, A., Farid, T., Farvid, M., Feigin, V., Ding, E. L., Fowkes, G., Gebrehiwot, T., Gillum, R., Gold, A., Gona, P., Gupta, R., Habtewold, T. D., Hafezi-Nejad, N., Hailu, T., Hailu, G. B., Hankey, G., Hassen, H. Y., Abate, K. H., Havmoeller, R., Hay, S. I., Horino, M., Hotez, P. J., Jacobsen, K., James, S., Javanbakht, M., Jeemon, P., John, D., Jonas, J., Kalkonde, Y., Karimkhani, C., Kasaeian, A., Khader, Y., Khan, A., Khang, Y.-H., Khera, S., Khoja, A. T., Khubchandani, J., Kim, D., Kolte, D., Kosen, S., Krohn, K. J., Kumar, G. A., Kwan, G. F., Lal, D. K., Larsson, A., Linn, S., Lopez, A., Lotufo, P. A., El Razek, H. M. A., Malekzadeh, R., Mazidi, M., Meier, T., Meles, K. G., Mensah, G., Meretoja, A., Mezgebe, H., Miller, T., Mirrakhimov, E., Mohammed, S., Moran, A. E., Musa, K. I., Narula, J., Neal, B., Ngalesoni, F., Nguyen, G., Obermeyer, C. M., Owolabi, M., Patton, G., Pedro, J., Qato, D., Qorbani, M., Rahimi, K., Rai, R. K., Rawaf, S., Ribeiro, A., Safiri, S., Salomon, J. A., Santos, I., Santric Milicevic, M., Sartorius, B., Schutte, A., Sepanlou, S., Shaikh, M. A., Shin, M.-J., Shishehbor, M., Shore, H., Silva, D. A. S., Sobngwi, E., Stranges, S., Swaminathan, S., Tabarés-Seisdedos, R., Tadele Atnafu, N., Tesfay, F., Thakur, J., Thrift, A., Topor-Madry, R., Truelsen, T., Tyrovolas, S., Ukwaja, K. N., Uthman, O., Vasankari, T., Vlassov, V., Vollset, S. E., Wakayo, T., Watkins, D., Weintraub, R., Werdecker, A., Westerman, R., Wiysonge, C. S., Wolfe, C., Workicho, A., Xu, G., Yano, Y., Yip, P., Yonemoto, N., Younis, M., Yu, C., Vos, T., Naghavi, M., and Murray, C. Global, regional, and national burden of cardiovascular diseases for 10 causes, 1990 to 2015. *Journal of the American College of Cardiology*, 70(1):1–25, 2017. ISSN 0735-1097. doi: https://doi.org/10.1016/j.jacc.2017.04.052. URL https://www.sciencedirect.com/science/article/pii/S0735109717372443.

SONY, M. R. K. Uci heart disease data. *Kaggle*, 2020. URL https://www.kaggle.com/datasets/redwankarimsony/heart-disease-data.

Xue, B., Zhang, M., and Browne, W. N. Particle swarm optimisation for feature selection in classification: Novel initialisation and updating mechanisms. *Applied Soft Computing*, 18:261–276, 2014. ISSN 1568-4946. doi: https://doi.org/10.1016/j.asoc.2013.09.018. URL https://www.sciencedirect.com/science/article/pii/S1568494613003128.