Hybrid Genetic Algorithm-SVM model for Heart Disease Classification

Mahmoud Bahaa, Hady Ragab, Karim Atif, Ezzat Ashraf, Ahmed Abdelaziz and Omar Magdy

mahmoud.b1996@gmail.com, Hadyragab74@gmail.com, karimatif21@gmail.com, ezzatashraf.1996@gmail.com, Ahmedbastawesi97@gmail.com, Omarmagdy718@gmail.com

Faculty of Computer Science- Ahram Canadian University, Cairo, Egypt
Supervisor: Mohammed M. Eissa, Assistant professor
Ahram Canadian University, Egypt, mohammed.mamdouh@gmail.com

Abstract– Heart disease has been known as one of the most reasons of mortality all around the world. By the by, this disease is preventable in the event that it can be analyzed in an early stage. In this manner, it is significant to create decision support systems that are able to assist doctors analyze the disease and its related dangers. Machine learning algorithms are commonly used in medical classification systems. Medical datasets are prone to having redundant features. Redundant features tend to decrease the accuracy of the classification. Thus, removing the redundant features is a must to gain a better accuracy. This paper presents a hybrid model for heart disease classification using genetic algorithm (GA) and support vector machine (SVM) techniques. The genetic algorithm is used in feature selection to optimize the dataset by eliminating the features that have no impact or decrease the accuracy. The SVM is used for classification on the optimized dataset. The data used in this paper is Cleveland heart disease dataset, taken from UCI machine learning repository. It consists of 303 cases with 13 features. The paper also presents a comparative analysis between the proposed model and another heart disease classification models. The experimental results show that, the proposed hybrid GA-SVM model have better accuracy than the other models.

Keywords-- Knowledge Discovery – Heart Disease – SVM – Genetic Algorithm – Rough Sets – Data Pre-processing. Classification – Cardiology – Artificial Intelligence– Diagnosis – Cardiovascular Disease – Diagnostic medicine– Data Mining.

I. INTRODUCTION

A. Problem Statement

Heart-related diseases or Cardiovascular Diseases (CVDs) are the main reason for a huge number of death re-phrase in the world over the last few decades and has emerged as the most life-threatening disease, not only in Egypt but in the whole world. It is the hardening of the blood vessels by fatty deposits called plaque [1]. The heart must get oxygen and nutrients to working well. Blood carries the oxygen and nutrients to the heart through the blood vessels called arteries. As the plaque builds up, blood flow to the heart muscle is decreased. When blood flow is decreased, it can cause chest pain, shortness of breath, or a heart attack to occur [2].

Today, data mining has grown so vast that they can be used in many applications; examples include predicting costs of corporate expense claims, in risk management, in financial analysis, in insurance, in process control in manufacturing, in healthcare, and in other fields [3].

This led to the use of knowledge discovery in databases in medical informatics, the database that is found in the hospitals, namely, the hospital information systems (HIS) containing massive amounts of information which includes patient's information, data from laboratories which keeps on growing year after year. With the help of knowledge discovery in databases methods, useful patterns of information can be found within the data, which will be utilized for further research and evaluation of reports. The other question that emerges is the manner by which to classify or assemble this enormous amount of data. The automatic classification technique is done dependent on similitudes present in the information. The Classification process is possibly demonstrated productive if the end that is drawn by the automatic classifier is acceptable to the cardiologist [4].

This paper presents a model based on such algorithms and techniques and analyzes the performance. Model-based on supervised learning algorithms such as Genetic Algorithm (GA), Support Vector Machine (SVM).

B. Support Vector Machine

A Support Vector Machine (SVM) is AI calculation that examines information for arrangement and relapse examination. SVM is a regulated learning technique that takes a gander at information and sorts it into one of two classifications. The SVM yields a guide of the arranged information with the edges between the two as far separated as conceivable it makes a level limit called a hyperplane, which isolates the space to make genuinely homogeneous parcels on either side. SVMs are utilized in content order, picture characterization, penmanship acknowledgment and in the science [5].

The target of the support vector machine algorithm is to discover a hyperplane in N-dimensional space (N — the number of features) that have the best classification process to classify data points. To isolate the two classes of data points, there are numerous conceivable hyperplanes that could be chosen. Our objective is to find a plane that has the maximum margin, the maximum distance between data points of both classes. Maximizing the margin distance provides some reinforcement so that future data points can be classified with more confidence [6].

C. Genetic Algorithm:

Genetic algorithm (GA) is an optimization technique based on the principles of natural selection. it's used to find the optimal or near-optimal solutions to tough problems that otherwise would take a very long time to solve. it's often used to solve optimisation issues in machine learning. The genetic algorithm consists of five phases as shown at fig (1) [7].

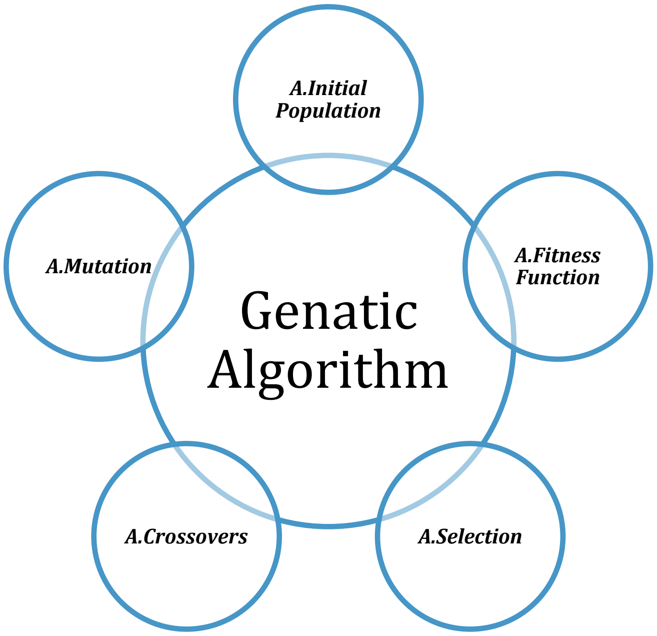

Fig.1 Genetic Algorithm Five Phases

1. *Initial Population*

Initial population is the first phase in the Genetic Algorithm process, a population is a set of chromosomes in the current generation. The initial population is the very first set of chromosomes which is usually random.

2. *Fitness Function*

The fitness function evaluates the fitness of an individual. A fitness score is given for each individual, the probability of being chosen in the selection phase is based on the fitness score.

3. *Selection*

Selection is the process of selecting pairs of individuals randomly to produce a new offspring. The higher the fitness score of an individual the higher their chance of being chosen.

4. *Crossovers*

Crossover is the process of producing an offspring, it is done by selecting a random point for each pair and exchanging genes between the two individuals.

5. *Mutation*

Like biological mutation, a mutation is random a change that occurs in one or more genes of an individual. The genes value change from its initial state to another state, but the chance of mutation is very small. The purpose of mutation is producing new solutions to the problem [7],[8].

II. Previous Work

Amid the most recent decade, an enormous measure of issues identifying with coronary heart disease (CHD) was examined. The treatment of CHD is a standout amongst the most significant issues. Numerous analysts have handled medical researchers in the area of treatment of CHD and new information appears frequently.

In 2011, B.K. Tripathi and et-al, A framework for intelligent medical diagnosis using rough set with formal concept analysis. In this study, the rule generation algorithm of rough set theory generates 91 rules. This is further minimized to 72 candidate rules with the help of domain intelligence and is further minimized to 65 rules by validation process and threshold value. Further these suitable rules are explored to identify the chief characteristics affecting the relationship between heart disease and its attributes by using formal concept analysis. This helps the decision maker a priori detection of the heart disease. Accuracy reached to 88% [9].

In 2012, S. Muthukaruppan ―A hybrid particle swarm optimization based fuzzy expert system for the diagnosis of coronary artery disease‖ have offered. Mamdani fuzzy inference system and have used the triangular membership functions, their proposed classification accuracy was 93.27 percent [10].

In 2015, V Krishnagar and his colleagues in their study entitled "Diagnosis of heart disease using fuzzy classification techniques" used a combination of fuzzy techniques and k-nearest neighbour, the proposed algorithm accuracy was 83.7% [11].

In 2018, Navdeep Singh and et-al, Heart Disease Prediction System using Hybrid Technique of Data Mining Algorithms. In this study, our aim was to design a heart disease prediction system using various data mining techniques and to perform the analysis of the results obtained for all implemented techniques. So, for the completion of the heart disease prediction model survey, we have evaluated the popular and effective heart disease prediction methods from the literature survey and finally select the most effective algorithms of Naïve Bayes and Genetic Algorithm for their performance analysis on the heart disease prediction. The performances of the models were evaluated using the genetic algorithm and naïve Bayes and achieved an accuracy of 97.14% [12].

In 2015, T. Santhanam his colleagues in their study entitled "Heart Disease Prediction Using Hybrid Genetic Fuzzy Model" used a combination of Genetic algorithm and Fuzzy techniques to create a system that aids the physicians for accurate prediction of heart disease in patients has been devised using the computing techniques like genetic algorithms and fuzzy logic and achieved an accuracy of 86% [13].

In 2018, Kathleen H. Miaoa and colleagues developed an enhanced deep neural network (DNN) learning to aid patients and healthcare professionals and to increase the accuracy and reliability of heart disease diagnosis and prognosis in patients and an accuracy of 83.67% was achieved [14].

III. Proposed Model

In this paper, Cleveland Heart Disease dataset is used obtained from the UCI repository [15]. The dataset contains large and continues values, so we used RSBR data pre-processing to convert continuous values to discrete and convert large values to small values. Feature selection with genetic algorithm is then used to remove irrelevant features to reach the optimal result. Now the dataset is ready for classification, since our dataset is non-linear, we used Support Vector Machine with kernel trick for classification, SVM works by maximizing the hyperplane between multiple classes. All of these steps shown at fig.2 [16].

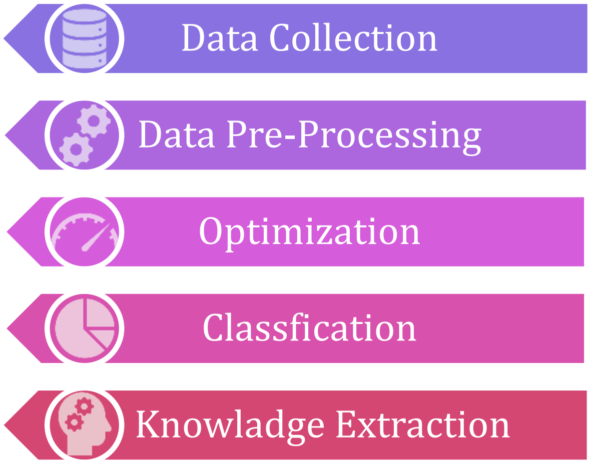

Fig.2 Proposed Model

A. Data Collection:

The Coronary Heart Disease dataset is obtained from the UCI repository. The Coronary Heart Disease dataset contains 13 attributes and 1 decision class. Most of attributes are numerical and there are only three attributes are binary attributes and all attributes have no missing value [15].

B. Data Pre-processing:

Data pre-processing is needed because the dataset contains continuous values. One method of pre-processing is RSBR discretization algorithm. Discretization is the process of transforming continuous values to discrete values. The goal of discretization is finding a set of cut points to shrink the ranges of values to the smallest number of intervals possible. These intervals are then used to find an equation and use Boolean algebra to simplify the equation as much as possible, from this equation the cut points can be found. RSBR Discretization Algorithm as shown at Table 1 [16].

C. Data Optimization:

Optimization is needed to gain a better accuracy, there are many techniques of optimization one of the most advanced techniques is feature selection using Genetic Algorithm (GA). Feature selection is the process of finding the most relevant variables for a predictive model. These techniques can be used to identify and remove unneeded, irrelevant and redundant features that do not contribute or decrease the accuracy of the predictive model. Genetic Algorithm optimization steps as shown at Table 2 [17].

TABLE I

RSBR DISCRETIZATION ALGORITHM

```
RSBR Discretization algorithm
Input
 Information system table ($S$) with real valued
          attributes $A_{ij}$
          And  $n$ is the number of inter values for each
          Attribute.
Output
          Information table ($ST$) with descretized real
          Valued attribute.

Begin procedure

    (1)     For $A_{ij} \in S$  do

    (2)     Define a set of Boolean variables as follows:

            $BV(U) = \{\sum_{i=1}^{n} C_{ai}, \sum_{i=1}^{n} C_{bi}, \sum_{i=1}^{n} C_{ci}, \dots \sum_{i=1}^{n} C_{ni}\}$

          Where $\sum_{i=1}^{n} C_{ai}$ corresponds to a set of intervals
          Defined on the variables of attributes a.

    (3)     End for

    (4)     Create new information table $T^P$ by using set of
        intervals

    (5)     Find minimal subset of  $C_{ai}$  that discerns all the
        objects in decision class D using the following
        formula:

            $\Phi^U = \wedge \{\psi\ (i,j)\ : d(x_i) \neq d(x_j)\}$

Where  $\psi\ (i,j)$  is the number of minimal cuts which must
be used to discern two different instances  $x_i$ and $x_j$ in the
information table.

End procedure
```

D. Classification:

There are many data mining techniques to classify data, but the support vector machine is one of the best techniques to classify nonlinear problems. A support vector machine (SVM) is machine learning algorithm that analyzes data for classification and regression analysis. SVM is a supervised learning method that looks at data and sorts it. An SVM outputs a map of the sorted data with the margins between the two as far apart as possible. SVM classification method shown at Table 3.

TABLE 2

GA OPTIMIZATON STEPS

1. Generate random population of *n* chromosomes (suitable solutions for the problem)
2. Evaluate the fitness *f(x)* of each chromosome *x* in the population
3. Create a new population by repeating following steps until the new population is complete
4. Select two parent chromosomes from a population according to their fitness (the better fitness, the bigger chance to be selected)
5. With a crossover probability cross over the parents to form a new offspring (children). If no crossover was performed, offspring is an exact copy of parents.
6. With a mutation probability mutate new offspring at each locus (position in chromosome).
7. Place new offspring in a new population
8. Use new generated population for a further run of algorithm
9. If the end condition is satisfied, **stop**, and return the best solution in current population
10. Go to step **2**

TABLE 3

SVM CLASSIFICATION MAP

1. Data pre-processing involves Dividing the data into attributes and labels and dividing the data into training and testing sets.

2. all the columns of the bank data frame are being stored in the X variable except the "Class" column, which is the label column. The drop () method drops this column.

3. In the second line, only the class column is being stored in the y variable. At this point of time X Variable contains attributes while y variable contains corresponding labels.

4. Once the data is divided into attributes and labels, the final pre-processing step is to divide data into training and test sets. Luckily, the model selection library of the Scikit-Learn library contains the train, test, split method that allows us to seamlessly divide data into training and test sets.

5. divided the data into training and testing sets.

6. Scikit-Learn contains the SVM library, which contains built-in classes for different SVM algorithms. Since we are going to perform a classification task, we will use the support vector classifier class, which is written as SVC in the Scikit-Learn's SVM library.

E. *Knowledge Extraction:*

Coronary heart disease is most commonly due to atherosclerotic occlusion of the coronary arteries. Atherosclerosis is a process that can involve many of the body's blood vessels with a variety of presentations. Half of all deaths in the developed world and a quarter of deaths in the developing world are due to Cardiovascular Disease which is comprised of hypertension and the diseases caused CHD [8].

By this paper, now we can extract knowledge from the dataset and can decide each patient may have or haven't coronary heart disease before doing Cardiac catheterization that will help The cardiologist to stands on the patient's condition because Cardiac catheterization has many risks on the patient health.

IV. Experiment Design

In this experiment the medical data related to Heart diseases is considered. This dataset was obtained from Cleveland database. This is publicly available dataset in the Internet. Cleveland dataset concerns classification of person into normal and abnormal person regarding heart diseases.

A. Dataset description

Used Dataset to test our model, we used the Cleveland Cleveland Heart Database taken from UCI learning data set repository. The dataset contains 13 numerical attributes and 1 decision attribute described in Tables 4 [7].

TABLE 4
Dataset description

Attribute	Description	Range
Age	Age in years	Continuous
Sex	(1=male; 0=female)	0,1
Cp	--Value 1: typical angina --Value 2: atypical anginal --Value 3: non-anginal pain --Value 4: asymptotic	1,2,3,4
trestbps	Resting blood pressure (in mm Hg)	Continuous
chol	Serum cholesterol in mg/dl	Continuous
fbs	(Fasting blood sugar .120mg/dl) (1=true; 0=false)	0,1
restecg	electrocardiography results --Value 0: normal --Value 1: having ST-T wave abnormality (T wave inversions and/or ST Elevation or depression of>0.05mV) --Value 2: showing probable or definite left ventricular Hypertrophy by Estes' criteria	0,1,2
Thalach	Maximum heart rate achieved	Continuous
Exang	Exercise induced angina(1=yes;0=no)	0,1
Old peak	ST depression induced by exercise relative to rest	Continuous
Slope	The slope of the peak exercise ST segment Value 1: up sloping Value 2: flat Value 3: down sloping	1,2,3
Ca	Number of major vessels (0-3) Coloured by fluoroscopy	Continuous
Thal	Normal, fixed defect, reversible defect	3,6,7
Target	0: Normal Person 1: Person with heart disease	0,1

B. Results

In this experiment a total of 303 records and 13 attributes (features) gathered from the UCI repository Cleveland Coronary Heart Disease data set the features were first optimised, the redundant and irrelevant features were removed which was done randomly, thus leaving us with an optimised dataset that is ready for classification. The records were split 80% training dataset and 20% testing dataset the training accuracy was 92% and the testing accuracy was 97% and the average is 95%.

C. Comparative study

In this paper, a Hybrid Genetic Algorithm-SVM model for to classify heart disease and predict with CHD treatment through the GA algorithm optimizing and doing feature selection by passing dataset into five phases of GA algorithm then SVM algorithm classify the features that been selected by GA algorithm to predict CHD treatment. A comparative analysis between the proposed model and other Heart Disease classification models then to determine proposed model efficiency.

In 2012 the paper "A Hybrid Particle Swarm Optimization Based Fuzzy Expert System for The Diagnosis of Heart Disease" was published, it used particle swarm and fuzzy logic and an accuracy of 93.27% was achieved [10]. Then in 2014 the paper "Diagnosis of heart disease using fuzzy classification techniques" was published it used fuzzy logic and an accuracy of 83.7% was achieved [11].

In 2015 the paper "Heart Disease Prediction Using Hybrid Genetic Fuzzy Model" was published it used Genetic algorithm and fuzzy logic and an accuracy of 86% was achieved [12]. Then in 2018 the paper "Coronary Heart Disease Diagnosis using Deep Neural Networks" was published it used deep neural network and an accuracy of 83.67% was achieved [13].

The Comparative study shows that the proposed model classification Accuracy is higher than the other models. Comparative results Shown at Fig3 and Table 5.

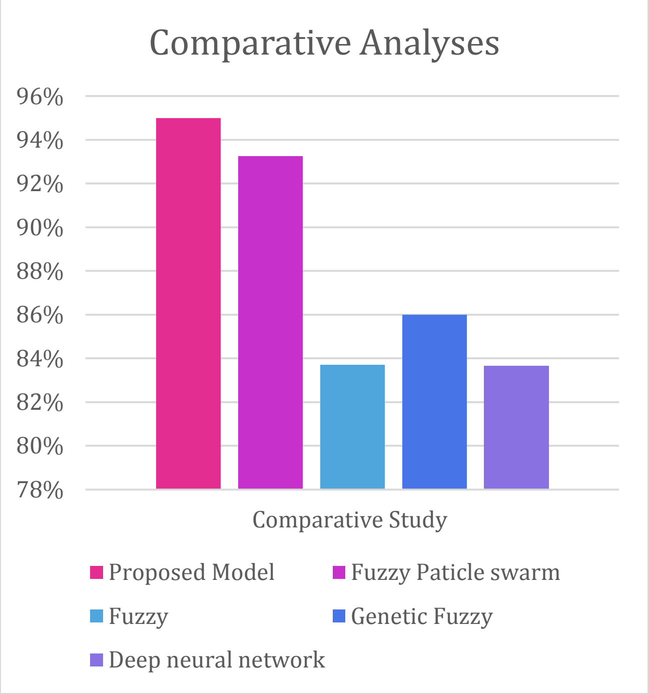

Fig.3 Comparative Analyses

TABLE 5
COMPARATIVE STUDY

Model	PM	FPS	Fuzzy	GA-F	DNN
Accuracy	**95%**	**93.27%**	**83.70%**	**86%**	**83.67%**

V. CONCLUSION & FUTURE WORK

Data Mining is a set of methodology that applies to giant and complicated databases. this can be to eliminate the randomness and see the hidden pattern. As these data processing ways square measure nearly always computationally intensive. we tend to use data processing tools, methodologies, and theories for revealing patterns in knowledge. Data mining can be used by medical experts to identify heart disease and help them make decisions. In this paper we used Genetic Algorithm and Support Vector Machine hybrid model for classification and achieved a high accuracy compared to different Heart Disease classification models. In the future, we are looking forward to using a larger dataset which may increase our accuracy.

ACKNOWLEDGMENT

Thanks for ALLAH for his generous blesses and gifts to us and for helping me to accomplish this Paper.
Many thanks to Dr. Mohammed Mamdouh Eissa for giving us the wonderful opportunity to complete this paper under his supervision, it is truly an honour. Thank you for all the advice, ideas, moral support and patience in guiding us through this project. Thank you for your enthusiasm for the encourage us. Your wealth of knowledge in the field of Data Science is inspiring.
Many thanks to Our parents and family for continuing to be part of our strong foundation, their constant support and never-ending encouragement.

References

[1]C. Debra Sullivan, "Heart disease: Types, causes, and treatments", Medical News Today, 2019. [Online]. Available: https://www.medicalnewstoday.com/articles/237191.php.

[2]J. Betancourt, A. Green, J. Carrillo and E. Park, "Cultural Competence And Health Care Disparities: Key Perspectives And Trends", Health Affairs, vol. 24, no. 2, pp. 499-505, 2005.

[3]V. Poornima and D. Gladis, "Analysis and Prediction of Heart Disease Aid of Various Data Mining Techniques: A Survey", International Journal of Business Intelligence and Data Mining, vol. 1, no. 1, p. 1, 2018.

[4] S. . Ishtake and S. . Sanap, "‘ Intelligent Heart Disease Prediction System Using Data Mining Techniques ’," International Journal of healthcare & biomedical Research, vol. 1, no. 3, pp. 94–101, 2013.

[5]AK, Suykens Johan. Least squares support vector machines. World Scientific, 2002.

[6]Babaoglu, İsmail, Oğuz Findik, and Erkan Ülker. "A comparison of feature selection models utilizing binary particle swarm optimization and genetic algorithm in determining coronary artery disease using support vector machine." Expert Systems with Applications 37.4 (2010): 3177-3183.

[7] D.E. Goldberg, Genetic Algorithms in Search, Optimization and Machine Learning, Addison Wesley, Reading, MA, 1989.

[8]Eissa M. M., Elmogy M., Hashem M., Badria, "Hybrid Rough Genetic algorithm Model for Making Treatment Decisions of Hepatitis C.", In. 2nd International Conference for Engineering and Technology ICET, GUC, Cairo, Egypt, 2014.

[9] Tripathy, B. K., D. P. Acharjya, and V. Cynthya. "A framework for intelligent medical diagnosis using rough set with formal concept analysis." arXiv preprint arXiv:1301.6011,2013.

[10] Muthukaruppan, S., and Meng Joo Er. "A hybrid particle swarm optimization based fuzzy expert system for the diagnosis of coronary artery disease." Expert Systems with Applications39.14 (2012): 11657-11665.

[11]Krishnaiah, V., et al. "Diagnosis of heart disease patients using fuzzy classification technique." International Conference on Computing and Communication Technologies. IEEE, 2014.

[12]Dewan, Ankita, and Meghna Sharma. "Prediction of heart disease using a hybrid technique in data mining classification." 2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom). IEEE, 2015.

[13] Santhanam, T., and E. P. Ephzibah. "Heart disease prediction using hybrid genetic fuzzy model." Indian Journal of Science and Technology 8.9 (2015): 797.

[14]Miaoa, Kathleen H., and Julia H. Miaoa. "Coronary Heart Disease Diagnosis using Deep Neural Networks." INTERNATIONAL JOURNAL

OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS 9.10 (2018): 1-8.

[15]"Heart Disease UCI", Kaggle.com, 2019. [Online]. Available: https://www.kaggle.com/ronitf/heart-disease-uci.

[16]Chao, S., Li, Y.: Multivariate interdependent discretization for continuous attribute. In: Proceeding of the 3rd International Conference on Information Technology and Applications, vol. 1, pp. 167–172 (2005).

[17] Rhodewalt & Smith," current issues in Type A behavior, coronary proneness, and coronary heart disease". In C.R. Snyder & D.R.Forsyth (Eds.), Handbook of social and clinical psychology New York: Pergamon,1991, pp.197-220.