Detection of Fraud and Corruption in Healthcare System

João Galamba de Olivera
Instituto de Telecomunicações
Instituto Superior Técnico
Lisbon, Portugal
joao.amorim.queiros.galamba.oliveira@ist.utl.pt

Abstract—This paper purposes a data mining technique for the detection of fraud and corruption in healthcare system. The developed genetic algorithm explores the relation between claims, patient and providers introducing a new approach for the combat of fraud and corruption in healthcare. The application of this method to the healthcare system will accelerate and improve the detections of both, well-known fraud cases and new fraudulent situations.

Keywords— healthcare; fraud; corruption; genetic algorithms; data mining

I. INTRODUCTION

With the increase of the average lifespan (in half century, average life spans increased 20 years [1]), mainly due to developments in medicine methods, healthcare system as grown both in its importance and in costs. In Portugal, the government spent 7.5 billion euros (5% gross domestic product) of the state budget in 2014. In the USA [2], annual expenditures approached the two trillion dollars in 2004, which corresponds to 15.3% of the gross domestic product. According to the European Anti-Fraud Organization, in 2009 were spend between 3% and 10% of the GDP in healthcare in the European countries [3] [4]. As an example, in 1999 coronary heart diseases cost £1.73 billion to the U.K. health system, £2.42 billion in informal care and £2.91 billion in productivity loss.

According to the Global Health Care Anti-Fraud Network [5], GHCAN, corruption increased as well. The National Health Care Anti-Fraud Association, NHCAA, in the United States, which works in association with GHCAN and EHFCN, among others, stands that about 9%, 47.9 billion dollars, in the of the United States annual healthcare expenditure, are lost due to fraud and corruption. The EHFCN stands that the corruption values in Europe round the 10% of the total investment.

Nowadays, the majority of the detection of fraud techniques in healthcare, rely on human experience to review insurance claims an identify the suspicious ones [6]. This process needs intensive medical knowledge and consumes a lot of time to analyze all situations. In healthcare fraud detection, there are three main entities involved: service providers (doctors, hospitals, ambulance companies and laboratories), insurances subscribers (patients and patients’ employers) and insurances carries (governmental health departments and private insurance companies). In order to classify the cases of fraud and corruption exists four categories, directly associated with the entities involved [7]: services provider’s fraud, insurance subscriber’s fraud, insurance carrier’s fraud and conspiracy fraud.

Typically, the cases of fraud happen when: home or hospital stay conflict, hospital stay with no associated physician inpatient visit, occurs excessive lab or radiology services per client per day, X-ray are duplicated, fragmented lab and X-ray procedures, x-ray interpretation with no associated technical portion and ambulance trips with no medical services [3] [8].

The principal contributions of this investigation are the clear organization of the healthcare data according to three main areas, claims’ information, patients’ information and providers’ information. The application of three independent genetic algorithms to each of the areas above described. And finally, the combination of the fraudulent classification criteria according to the results of the independent genetic algorithms.

The paper is organized in five sections: The first, Introduction, explains the Healthcare System, the different anti-fraud organizations and the corruption structure. The second, Related Work, describes the Portuguese healthcare situation, the healthcare data process and machine learning techniques to detect cases of fraud as well as specific techniques that are used. The third, Architecture, describes the GA structure and it’s specific functions. The fourth, Analysis of Results, examine the algorithm performance and presents the different simulations. The last, Conclusions, interpreters the results of the tests presented in the previous section, as well eventual changes of the algorithm structure.

II. RELATED WORK

A. Portuguese Situation

In Portugal the most common corruptions cases are related in one way with informal payments in medical service delivery and, in another perspective, when doctors receive presents or sponsorship to participate in congresses, or other events, by some pharmaceutical representative. In the last 10 years were registered four big fraud cases [6]: 2010 - “Remédio Santo”, 2011 - “Castle”, 2012 - “Remédio Jovem”, and 2013 - “Recife de Saúde”.

The EHAN (European Health Anti-Fraud Network) and the European Anti-Fraud Organization (EHFCN) were created and the activities are related to the prevention and detection of fraud and corruption in health care. The EHAN promotes the cooperation of anti-fraud agencies in the European Union and the EHFCN monitors the occurrence of fraud and corruption in healthcare system of European Union.

B. Fraud Detection

Typically, the detection of frauds is a difficult and time consuming task. The proposed method uses machine learning techniques to extract features in healthcare data, and then, the genetic algorithm classifies the cases of fraud. The algorithm is composed of four parts: the first, Related Work, describes the process of data mining and genetic algorithms in fraud detection. The second, Architecture, explains the GA structure and its specific functions. The third, Related Work, describes the process of data mining and genetic algorithms in fraud detection. The fourth, Analysis of Results, presents the results of the proposed method. The last, Conclusions, interpreters the results of the tests presented in the previous section, as well eventual changes of the algorithm structure.

C. Genetic Algorithms

The genetic algorithm (GA) is a search technique used in computing to find the solution to complex optimization problems. The algorithm is inspired by the process of natural selection and evolution. It begins with a population of candidate solutions and progresses towards better solutions by using techniques such as selection, crossover, and mutation. Genetic algorithms are often used in machine learning for feature selection, parameter optimization, and other tasks.

D. Fraud Detection in Healthcare

The detection of fraud in healthcare systems is a complex task due to the large amount of data and diversity of fraud types. Fraud detection techniques can be classified as rule-based, data mining, and machine learning methods. Rule-based methods use predefined rules or patterns to identify fraudulent behavior, while data mining techniques analyze data patterns and relationships to detect anomalies. Machine learning methods, including genetic algorithms, learn from data to identify fraudulent behavior.

E. Machine Learning Techniques

Machine learning techniques, such as neural networks, support vector machines, and decision trees, are used in healthcare fraud detection to identify patterns and anomalies in the data. These methods can be used to classify fraudulent behavior into different categories, such as fraud by providers, patients, or carriers. Genetic algorithms can be used to optimize the selection of features and parameters for these machine learning techniques.

F. Genetic Algorithms in Healthcare Fraud Detection

The genetic algorithm is a population-based optimization method that mimics the process of natural selection and evolution. It is used in healthcare fraud detection to optimize the selection of features and parameters for machine learning techniques. The genetic algorithm can improve the performance of these techniques by identifying the best combination of features and parameters that maximize the detection of fraudulent behavior.

G. Conclusion

The proposed method uses machine learning techniques and genetic algorithms to detect fraud in healthcare systems. The genetic algorithm is used to identify the best combination of features and parameters for the machine learning techniques. The results of the proposed method show that the combination of genetic algorithms and machine learning techniques can improve the detection of fraudulent behavior in healthcare systems.
Healthcare systems in Portugal present four main areas with weakness, according to the European Union entities [6]: 1 - Weak relation between the public sector and the private services entities. 2 - The restructuring of the internal management system. 3 - Medicines are purchased on wholesale, instead of unit sale. 4 - Weak control mechanisms and complicated legal framework in corruption issues.

### B. Data Evolution

Processing the data available is of major importance. The following method describes a systematic procedure used to prepare the data [3]. 1 – Data Acquisition: Obtain a reliable and detailed data base; 2 – Goal Setting: Identify and prioritize the types of fraud on which detection should be focused; 3 - Data cleaning: Analyses of the data, identification and correction of similar information described in different forms, that may create new variable without need; 4 - Handing missing values: Application of attribution of information throw the hot-deck imputation method, based on the replacement of the value with the most similar one, or with regression imputation, based on the prediction of the value throw the predicted value of the regression. 5 - Data transformation: Adaptation of the data for the pretended objective and respective algorithm; 6 - Feature selection: Application of the algorithm; 7 - Data auditing: Statistical check and visualization to become familiar with the data;

### C. Typologies of Corruption

The typologies numerated in the following, represent a classification of fraud cases [6].

1. Patient and Providers Fraud;

**Bribery in Medical Delivery:** In this fraud brides are paid in order to receive special care from providers. Many cases are related to procurement and purchase of pharmaceutical and medical services. Normally the brides are made in cash and it could be offers from the patient or even demanded by the service provider.

**Excessive number of clinical trials:** This fraud relates the number of clinical trials with the real necessity of this medical interventions. In this typology providers suggest an unexpected high number of clinical trials, which were not necessary, increasing the expenses supported by the patient unnecessarily.

**Private and Public services relation:** This typology of fraud refers to providers that work in both the public and in the private healthcare services. In this scenario the provider takes advantage of a patient from the public healthcare system and redirects the patient to the private healthcare system.

2. Payers and Providers Fraud;

**Undue Reimbursement Claims:** In this scenario, the health care providers recommend a medicine or a treatment, which is complex to define, and therefore, the patient does not have a clear notion of the importance of that situation. Generally known as Providers Payment Mechanism, PPM, which are rules and procedures for filling a claim. In these situations, the provider should present in a clear way the subject, which normally end in a very detail and sophisticated descriptions of the service and becomes useless for the patient to decide what is better. In some situation the insurance may not guarantee a specific treatment and the patient does not even know of that.

**Unexpected High Costs:** In this typology providers claim unexpected high costs for treatments that are significantly higher than the average value of costs for that claim. Normally providers take advantage of the lack of healthcare understanding from the patient increasing the total costs of the medical situation, due to extra services provided.

**Ghost Claims:** In this fraud, providers declare a claim, however there was no patient in the real claim. Both provider and patient are involved in this typology of fraud. The provider, because simulates the medical situation, and the patient, because is in conform the report of the medical provider. The objective is to obtain income from the insurance companies that cannot control effectively of the claim really took place.

3. Providers and Industry Fraud:

**Procurement Corruption:** This typology is related with pharmaceutical industries and medical device companies when some contract is performed with the healthcare providers. In this scenario brides can be offered to individual agents, such as money, leisure and trips and favoring relatives, or, in another perspective, to medical institutions, in which brides can be money, conference participation, free supply of materials and other kind of nonmonetary gifts.

**Mislese of (high) level positions:** This typology connects high-level political parties and administrative positions that are actively related to corruption cases. These entities are normally directly involved with the industry or healthcare system. There is not always a direct link between the corrupt interaction and the desired outcome, which makes these cases extremely difficult to proof. Regulators, political parties, industry and healthcare providers are the principal entities involved in this typology.

**Fraud and embezzlement of medicines and medical devices:** This typology is related with health care providers, in this case doctors, pharmacists and other technical personal, which have special access to medicine and medical facilities that are meant to be use. However, instead of being used with the initial propose, they are exposed as private business or sold overseas in health system with other regulations.

4. Industry and Regulators Fraud:

**Improper Marketing Relations:** This corruption cases involve the pharmaceutical and medical device industries with the healthcare providers and healthcare regulators. The industries influence the healthcare entities in order to obtain a bigger use of their products. The money from pharmaceutical companies and the influence it buys is in integral to the way the healthcare sector works. In one hand, industries collaborate with doctors to finance new investigations. In the other hand,
healthcare providers promote their products.

D. Machine Learning Techniques

Nowadays, the detection of fraud uses both the Statistics approach and the Machine Learning methods to combat fraud. The uses of these two methods are commonly called of Data Mining in each are explored the artificial intelligence attributes to understand data and then they are compared using a statistic method. However, Data Mining is becoming a multidisciplinary field, which combines not only statistics and machine learning techniques, but also visualization, information science and database technology.

The methods to detect fraud and corruption in healthcare have three different classifications. Supervised methods, used when historical data is available, Unsupervised methods, used when there is no historical data, and Hybrid method, which combines both the Supervised and the Unsupervised methods. Usually Hybrid methods use Unsupervised methods to improve the performance of the Supervised method.

1) Supervised Methods

a) Neural Networks

Under the machine-learning methods, there is Artificial Neural Networks [2] [3]. It is inspired on the animals’ biological nervous system, the brain, when it processes information. It is mainly composed by highly connected elements, just like neurons, that work together to solve the problem. Neural Networks are especially useful because they can handle big amount of data and complex structures with no linear relationships.

Some of the nodes receive scalar data from other nodes and transform the information to a single output signal. The interconnections are weighted and the weights are tunable. This method is clearly useful for detection of fraud in healthcare, due to its enormous variety and quantity of data. This method also tolerates noisy data well.

b) Decision Trees

Another machine learning technique is using decision tree to identify services providers’ fraud, detecting insurance subscribers’ fraud and for planning audit strategies in fraud detection [2] [9]. Decision trees maps information by several classifiers following politics such as minimizing false positives, minimizing false negatives and achieve a tradeoff between false positives and false negatives. This method produces a set of classifiers and votes on them to classify cases.

Decision trees are a good approach to detect fraud in healthcare because of its simplicity in the interpretation of the results and its ability to both generate rules from the tree and handling missing values. However, exists too many rules generated for large dimensional databases and few adjustable parameters available.

c) Combined GA and KNN

This combined method uses Genetic Algorithm to discover the optimal weighting of the feature used to classify general practitioners practice profiles and then uses them in the KNN algorithm to identify the nearest neighbor practice profile [7]. To determine the practice profile is used both the majority rule and the Bayesian rule. The genetic algorithm improves the efficiency of the KNN algorithm by finding in a very efficient way the near optimal set of weights.

2) Unsupervised Methods

a) Self-Organizing Map (SOM)

A SOM is a derivation of Neural Network, however is trained using an unsupervised learning method [4]. Is applied to general practitioner’s database to create unbiased subdivision of general practitioners’ practices for a more effective monitoring test ordering. SOM provides a topology preserving mapping from the high dimensional space to map units. Points in the map that are near to each other, in the input space, are mapped to nearby map units in the SOM.

3) Hybrid Methods

a) SOM and Neural Networks

The principals of the two methods are identical [4]. After some experimental results, the conclusions were that both methods had good performances even working in different ways. The idea of this method is to combine the Supervised and the Unsupervised in a neural networks approach. The training data is initially divided into four classes indicating different possibilities of fraud. After the application of the Neural Network method, SOM is employed to refine the training data. The classifications uses both, the ones obtained by the SOM method and the ones given by domain experts.

E. Technical Adjustments

a) Clinical Path

This concept is important and normally receives a special attention from managers in large hospital. Clinical path is defined as multidisciplinary care plans, in which diagnosis and therapeutic intervention are performed by physicians, nurses and other staff for a particular diagnosis or procedure [10]. Clinical path agents are physician orders, clinical industry and local standards of care.

The clinical path helps the algorithm to take decisions and select the right rules to apply. This approach increases efficiency and enables best practice, by reducing rework and resource waste. In another hand, a care activity is very likely to be fraudulent if its sequence is suspicious. Clinical path can also be used as a filter to eliminate redundant and irrelevant information.

b) Geo-Location Clustering

This technique purposes a clustering model considering the geo-location information of both the beneficiary and the provider to identify cases of fraud [4]. When a beneficiary travels a long distance to obtain healthcare services this may imply fraud. Although long travel distances may happen when
the quality of service is better in a certain area or when a specific service is not provided in the area.

The models applied, however, do not take into account the type of fraud. Cases such us the use of other’s person beneficiary identification card or schemes of agreements between providers and beneficiaries to take advantage of such way are not considered in this method.

III. ARCHITECTURE

The algorithm approach that will be used to detect fraud and abuse in healthcare system is, as mention before, a Genetic Algorithm [11]. This is an evolutionary algorithm, which aims to obtain better solutions in a short time period. The objective of this model is to detect whether any fraud has occurred or not. The Figure 1 presents the simple scheme of the architecture of this algorithm.

![Figure 1: Simplified Algorithm Architecture](image)

A. Data Generator

It is very hard to have access to real world Healthcare data. The confidentiality of the relevant data is considerable important and both, the governmental and private institutions, that were contacted for the development of this research, did not provide any information for the test of the developed algorithm. While the negotiations with the governmental healthcare institutions were progressing, the need of a data-base to test and upgrade the algorithm forced the creation of a Data Generator based on a research of the available information [7] [12] [13]. The Data Generator is composed by information’s taken from claims.

Patient information:
- Patient ID: Identifies the patient with a single identification number in order to maintain his confidentiality;
- Patient Residence: In a range from 0 to 5, this field classifies the distance from the patient residence until the hospital where the claim took place;
- Patient Age;
- Patient Gender: Male or Female;
- Source of payment: Classifies the type of payment performed by the patient. The type of payment is organized according to: A (Government program); B (Insurance mechanism); C (Self-pay); D (No charge [free, charity, special research, teaching]); E (Other)

Provider Information:
- Provider ID: Identifies the provider with a single identification number in order to maintain his confidentiality;
- Provider Residence: In a range from 0 to 5, this field classifies the distance from the provider residence until the hospital where the claim took place;

Claim Information:
- Provider Information:
  - Patient ID: Value corresponds to the costs of the claim supported by the patient;
  - Insurance Cost: Value corresponds to the costs of the claim supported by the insurance company that sponsors the patient;
  - Patient Risk In: In a range from P1 to P5, this field classifies the risk of life of the patient immediately before the claim takes place. The risk classification is organized according to: P1 - Normal healthy patient. P2 - Patients with mild systemic disease. P3 - Patients with mild systemic diseases. P4 - Patient with severe systemic diseases that is a constant threat to life. P5 - Moribund patients who are not expected to survive without the operation) [14];
  - Patient Risk Out: In a range from P1 to P5, this field classifies the risk of life of the patient after the claim took place. Uses the same classification presented in the Patient Risk In;
  - Medicine ID: Identifies the medicine administrated to the patient;
  - Medicine Cost: Information about the total price of the medicines;
  - Medicine Quantity: Information about the quantity of medicines;
  - DRG Code: Diagnosis Related Group, provides information about the classification of the hospital case. The specific DRG codes used are explain in the paragraph ahead [15];
  - Hospital: Information about the institutions where the claim took place;
  - Service Charged: Costs related with treatments or analyses performed in the claim;
  - Disposition: Information about the continuity of treatment. It’s the plan for continuing health care of a patient.
  - Date: Information about the date and hour of the end of the claim;

The objective of the data generator is to simulate real healthcare situation, for that purpose, the data must be as real as possible. However, healthcare information is too large and detail to explore with a data generator, so was define that were going to be explored only four hospital cases, distinguished by the respective DRG code. C01 refers to major pulmonary procedures, C02 to major cardiac procedures, C03 stomach and gastro procedures, C04 kidney and urinary tract procedures.

B. Organization of Data

Before the application of the genetic algorithm the data must be organized and correlated. The objective of this section is to adapt the information and improve the algorithm performance in a pre-stage. The data organization is divided in four steps: The first, goal setting, is the detection of fraud. The second, data cleaning, will filer only the information that is useful for the determine objective. In this case, the data is generated so all the variables were simulated with a specific purpose and data cleaning is not a direct priority. Third, handling missing values, once again this was not a priority because the data is simulated. And at last, data transformation. In this section the claims are
grouped in three areas, individual claims, patient claims and provider claims.

The patient claims and the provider claims are the aggregation of the claims by each patient and provider, respectively. This approaches enables the correlation of information, such as number of claims per patient/provider, which increases the number of variables to explore. The final objective is to apply the genetic algorithm to these three areas and study the correlation of the cases that are considered fraud. This approach will reduce the number false positive alarms.

C. Genetic Algorithm Description

The GA combines the individuals/chromosomes of an initial population in order to obtain individuals with better solutions in the end. Each individual represents a criteria of evaluation of claims and will classify the claim as fraudulent or legitimate. The ideal individual will classify correctly all the fraudulent situations, as well as non-fraudulent all the normal situations. The application of this GA is divided in seven steps in order detect situations of fraud and abuse effectively [16]:

1. The initial population is selected randomly from the sample space, which has many populations;
2. A fitness value is calculated for each chromosome in each population;
3. Selection process, where two parent chromosomes are selected through the tournament method;
4. The Crossover forms new offspring from the parent chromosomes using single point probability;
5. Mutation is applied in the new offspring using uniform probability measure;
6. The elitism selection chooses the best solution and passed it to the further generation;
7. The new population is generated and undergoes the same process until the reaches the validation criteria;

The flowchart, in Figure 1, describes the simple GA architecture.

The chromosomes of the population used in each genetic algorithm will be classified according to their precision on detection the fraudulent situation, and change throw different methods in order to improve their results.

2) Operations

a) Crossover

The step Crossover is used to produce new chromosomes from a pair of encoded chromosomes, usual called of parents, and combining them to produce two different chromosomes, usual called progeny [11]. The two new chromosomes have the best attributes of its parents.

The genes of each parent are switch if the according to a probability of 70%. In the following is the Crossover pseudo code:

```
crossover (Parent X, Parent Y, indpb)

For i in range(size):
    if random() < indpb:
        Parent X[i] = Parent Y[i]
        Parent Y[i] = Parent X[i]
```

The chromosomes of the population used in each genetic algorithm will be classified according to their precision on detection the fraudulent situation, and change throw different methods in order to improve their results.

Table 1: Chromosome Description in the GAs

<table>
<thead>
<tr>
<th>First Algorithm</th>
<th>Second Algorithm</th>
<th>Third Algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>W. Patient Residence</td>
<td>W. Patient Residence</td>
<td>W. Provider Residence</td>
</tr>
<tr>
<td>W. Age</td>
<td>W. Age</td>
<td>W. Number of Claims</td>
</tr>
<tr>
<td>W. Gender</td>
<td>W. Gender</td>
<td>W. Patient Residence</td>
</tr>
<tr>
<td>W. Source of payment</td>
<td>W. Number of Claims</td>
<td>W. Patient Age</td>
</tr>
<tr>
<td>W. Patient Costs</td>
<td>W. Source of Payment</td>
<td>W. Patient Gender</td>
</tr>
<tr>
<td>W. Insurance Costs</td>
<td>W. Patient Costs</td>
<td>W. Source of Payment</td>
</tr>
<tr>
<td>W. Patient Risk In</td>
<td>W. Insurance Costs</td>
<td>W. Patient Costs</td>
</tr>
<tr>
<td>W. Patient Risk Out</td>
<td>W. Patient Risk In</td>
<td>W. Insurance Costs</td>
</tr>
<tr>
<td>W. Provider Residence</td>
<td>W. Patient Risk Out</td>
<td>W. Patient Risk In</td>
</tr>
<tr>
<td>W. Medicine ID</td>
<td>W. Medicine ID</td>
<td>W. Patient Risk Out</td>
</tr>
<tr>
<td>W. Medicine Costs</td>
<td>W. Medicine Costs</td>
<td>W. Medicine ID</td>
</tr>
<tr>
<td>W. Medicine Quantity</td>
<td>W. Medicine Quantity</td>
<td>W. Medicine Costs</td>
</tr>
<tr>
<td>W. Service Charged</td>
<td>W. DRG</td>
<td>W. Medicine Quantity</td>
</tr>
<tr>
<td>W. Date</td>
<td>W. Hospital</td>
<td>W. DRG</td>
</tr>
<tr>
<td>Criteria</td>
<td>W. Service Charged</td>
<td>W. Hospital</td>
</tr>
<tr>
<td>W. Date</td>
<td>W. Service Charged</td>
<td></td>
</tr>
</tbody>
</table>

Criteria

Figure 2: Flowchart of the Genetic Algorithm

1) Chromosome Description

The chromosome for the genetic algorithm is composed by genes or features, each associated with a specific area of information. The features are weights that describe the importance of the variable associated according to the type of fraud that the algorithm is trying to detect. The sum of all the weights of a chromosome is one, with the exception of the last gene which is an integer number that represents the criteria of validation, explained in section Fitness Function.

The Table 1 specifies the genes of the chromosomes in the three genetic algorithm. The first algorithm, referent to the claims, the second algorithm, referent to the patients and the last algorithm is focus on providers.
b) Mutation

The step Mutation is commonly defined as a minimal change to a chromosome [11] [17]. The two chromosomes are perturbed in small ways upward and downward. The weight value for each feature has a certain probability of being changed. This minimal change is performed normally in the end of each generation, right before the combination of the new chromosomes and the population are again combined.

The mutation process occurs with a probability of 20%. In the following is the Mutation pseudo code:

```
Mutation (Chromosome, generation)

For i in len(Chromosome)
    Value_mutation = random_value {gauss (0.2, 0.1)}
    If generation > 15:
        Value_mutation = random_value {gauss (0.3, 0.3)}
    If random () < 0.5:
        Chromosome [i] = Chromosome [i] * (1 + Value_mutation)
    Else:
        Chromosome [i] = Chromosome[i] * (1 + Value_mutation)
```

c) Additional Approaches

Besides the standard structure of the genetic algorithm were add four strategies to increase the results and accelerate the performance of the algorithm. The first is related with the worst chromosomes and the best chromosomes. In each generation the two chromosomes with lowest fitness will be removed and the two best will be duplicated and passed to the next generation.

The second strategy applies only on the five chromosomes with lowest fitness in the end of each generation. The algorithm will apply an extra mutation or recreate to the chromosome from the beginning with a probability of 60% and 40% respectively.

The third strategy applies on the three chromosomes with best fitness. These chromosomes will pass in the mutation process once more, in order to improve their fitness.

The fourth strategy is only applied when the algorithm takes too much time to find a chromosome with a good fitness. If the maximum fitness is below 0.15, then all the population is submitted to a process where all the genes have a 50% chance of being randomly recreated

D. Validation of the Last Generation

One important part in the application of GA is the termination rule. This termination rule may depend on the type of fraud and in the specificities of the detection method. In this case, the termination condition is composed by two possible situations, and its activated if one of them is achieved. The first, is achieve if the population of chromosome present a standard deviation lower than 0.005. The second is more complex, it involves three parameters: standard deviation, maximum value and minimum value of the fitness. To activate the termination, condition the standard deviation must be lower than 0.025, the maximum fitness of this generation must be equal to the previous eight generations and the difference between the minimum fitness of this generation and the previous five generations must not be higher than 10%.

E. Fitness Function

The fitness function will classify each chromosome according to the results of a linear equations that distinguish the cases that are considered fraud from the others. The fitness is divided in two parts, first the application of the linear equations and second, an evaluation of the veracity of the result. The linear equation, in equation 1, will multiply each gene per the respective variable associated. Being N the number of genes that forms the chromosome. As explain in section, Chromosome Description, the last feature, that represents the fraud criteria, is excluded from the equation below.

```
\sum_{i=0}^{N-1} Weight of the Chromosome[i] * Claim Information[i] \quad (1)
```

In the following are described the five rules implemented for the evaluation of specific information’s of the claims. Initially is presented the theoretical approach, followed by the main objective of that specific rule:

A – Mean Rule: Compares the claim value with the mean value of the population. If the claim value is 1.5, 2, 3 or 4 times higher than the mean value, the claim level is increased to level 1, 2, 6 and 10 respectively. Objective: Detect unexpected changes in the sample.

B – Max/Min variation Rule: If the difference between the maximum value, in the sample, and the mean value is two times higher than the absolute difference between the minimum and the mean, the claim value level increases 50%. Objective: Increase the importance of a specific variable in the case that the sample is not uniform.

C – Majority Rule: If the variable is not the major (i.e. used at least more 10% than the respective others in the sample), the claim is increases two levels. Objective: Explore incoherence in the variable administrated in comparison with the regular proceeding.

D – Linear Importance Rule: The variable observed by this rule are matrix of five positions. Each position is multiply by a differentiator, respectively 1, 2, 3, 4 and 5, followed by the sum of all the new values in the matrix position. This value will be the associated level of this variable. Objective: Classify the variable according to its course.

E – Regularity Rule: Analyses the standard deviation and the mean value of the matrix. If the ratio between the standard deviation and the mean value is less than one, the level of the variable claim increases 5 levels. Objective: Classify the regularity of the actions described by this variable.

After the application of the linear equations, the result is compared with the gene criteria. If the result of the claim is, according to the chromosome, equal or higher than the criteria, the claim is considered fraudulent. Each chromosome will classify all the claims according to its criteria and the result will determine if the claim is fraudulent or not.
Finally, the fitness function, in equation 2, will investigate the evaluations of the chromosome and determine the false positive and the true positive in all the data base. False positive, are the number of claims classified as fraudulent but are legitimate, divided by the total number of legitimate claims. True positive are the number of claims correctly classified as fraudulent, divided by the total number of fraudulent claims.

\[
\text{Fitness} = \frac{\text{True Positive}}{\text{False Positive}}
\]

(2)

F. Chromosome Performance

After the validation of the last generation, is selected the best chromosomes of each GA. The criteria associated with each chromosome is applied to the test sample. The test sample is three times smaller than the training sample, however the number of fraudulent claims is the same in both samples, as it is discriminated in section Analysis of Results.

The study is now focus on the intersection of the detection of fraudulent situations suggested by the three GA. In some situations, the criteria in the chromosomes of a specific GA may consider incorrectly a situation as fraudulent. This situation is avoided if the remaining algorithms do not conclude the same. The results of the GA Claims will be intersected with the results of the GA Patient and with the GA Providers. The intersections, A – B, A – C, B – C and A – B – C, being GA Claims, A, GA Patient, B, and GA Providers, C, will in some situations present better solutions that the individual GA.

Figure 1 presents a flowchart with the total algorithm structure, described previously.

Figure 3: Flowchart of the Algorithm Architecture

IV. ANALYSIS OF RESULTS

A. Types of Fraud Simulated

The generator of claims will create a training sample and a test sample, both with legitimate and fraudulent claims. The nine fraudulent scenarios are the following [6] [8] [18]:

1. **Unexpected value of costs (ValCost)** – This scenario are created claims where the value of the costs is between two and three times higher than the maximum possible for the normal generated situations. The three cost parameters, patient cost, insurance cost, serviced charged may be modified.

2. **High number of claims (NumClaim)** – This scenario is related with considerable superior number of claims taking by both the patient and the providers.

3. **Irregular method of payment (MethPay)** – In the perspective of the patient the method of payment is normally regular, i.e., for identical claims irregular or uncommon methods of payment are considered suspected. In the provider point of view, a frequent unusual method of payment is also considered suspicious.

4. **Unusual development of the patient risk (PatRisk)** – This scenario is focus in one hand in the progress of the patient risk, immediately before the claim takes place and immediately after it ends, and in another hand, with frequency of this cases. If high variances of the patient risk, from risk level P1 to risk level P4/P5, and the opposite, the claim is considered suspicious. Frequent high risk situations, risk level P4 or P5, are also considered suspicious.

4. **Unusual development of the provider risk (ProRisk)** - Similar to the previous fraud case, this fraud simulates suspicious changes of patient risk, from risk level P1 to risk level P4/P5, and the opposite in a specific provider. High risk situations, risk level P4 or P5, are considered suspicious, as well as high variances of the patient risk.

6. **Uncommon use of a specific medicine (UncMed)** – This fraud tests frequent use of unusual medicine in a normal situation. This scenario indicates suspicious preference of a certain type of medicine.

7. **Excessive use of a specific medicine (ExcMed)** – This scenario explores elevated values of prescript medicines. The quantities of medicine in this fraud are between two and four times higher than the maximum possible for the normal generated situations.

8. **Unexpected value of medicine costs (MedCost)** – This fraud situation describes anomalies in the expected price of a medicine. The quantities of medicine in this fraud are between two and four times higher than the maximum possible for the normal generated situations.

9. **Irregular medical distance (MedDist)** – This scenario is related with both the patient and the provider distance between the respective residence and the associated local where the claim took place. In this scenario a specific patient or provider receives frequent visits from a provider or patient that lives far from the local where the claim took place. Both level 4 and 5 of the range of the patient and provider residences are considered uncommon. A high quantity of this cases is considered suspicious in this scenario.

B. Results

1) **Individual Simulation**

The individual simulations were developed with a training population of 765 claims, in which, 15 are fraudulent claims. And a test population of 265 with 15 fraudulent situations.
a weight of 0.91. And the worse situation in simulation 2 and simulation 8, NumClaim and MedCost.

The relation between the performance of the criteria’s, developed in the GA in the training sample and tested in the test sample, is observed with the delta weigh.

Excluding simulation 1, ValCost, in which the test results are considerable worse than the training ones, and simulation 8, MedCost, in which the difference of weights is 0.5, the results are consistent, and the both the training and test samples have similar results.

2) Group Simulation Part I

These group simulations were developed with a training population of 765 with 15 fraudulent claims, and a test population of 265 with 15 fraudulent claims.

Table 4: Average Results of the first population in Group Simulation

| FT 1 – ValCost ; FT 2 – NumClaim ; FT 3 – MethPay ; FT 4 – PatRisk |
| FT 5 – ProRisk ; FT 6 – UncMed ; FT 7 – ExcMed ; FT 8 – MedCost |
| FT 9 – MedDist |
| FT 10 – FT 13 ; FT 14 |
| FT 15 – FT 18 |

The individual test of the fraud cases simulated in Table 2 and Table 3, had, a positive training performance with low number of false detection and a high number true detection. The best situations are observed in the simulations 1 and 6, ValCost and UncMed respectively, both with a weight of 0.98.

In the simulation 1, all the GA used are extremely sensitive to changes in the costs of the claims, which explains the good results, almost total detection, obtained with the training sample. This situation is also observed in simulation 6, where is tested fraud type 6, UncMed, that mainly influences the medicine used. In this cases, the three GA consider and compare the high use of a specific medicine, which enabled a good training performance, above 0.9 of weight.

The worse simulated case scenario is the simulation 2, NumClaim, with just 0.46 of weight. The number of claims of a specific patient is neither detected by the claims GA, neither by the provider GA. The performance of the second GA, dedicated exclusively to the patient perspective, as an important influence in the detection of this fraud case, considering that the number of claims variable is directly evaluated by the genes of the chromosomes in this GA.

With the test sample the scenario does not change. The best situation is observed as expected in simulation 6, UncMed, with

From Table 4, comparing with the results obtained in the individual simulations of fraud, in the section above, weights value reduces in both the training sample and test sample. The best training situation is presented in the simulation 14, with fraud type respectively, ValCost, ExcMed and MedCost. All the training examples have a low false positive rate, below 0.1 value considerably good which slightly increases in the test set to a maximum of 0.23. The true positive scenario is different, in the training sample of simulation 11 its just 0.62, and in the test sample it reduces to 0.23, keeping the tendency.

Besides the fact that this are average results of all the experiences, GA tests and intersections of the results. The weight value in test sample is clearly low, 0.1 in simulation 11 scenario, and delta weight is considerable high, which specifies the differences between the train and the test samples.

3) Group Simulation Part II

These group simulations were developed with a training population of 7550 with 50 fraudulent claims, and a test population of 2550 with 50 fraudulent claims.

Table 2: Average Results of the Individual Simulation (part: I)

| FT 1 – FT 4 |
| FT 5 – FT 8 |
| FT 9 |
| FT 10 – FT 13 |
| FT 14 |

The relation between the performance of the criteria’s, in simulation 1, ValCost, in which the test results are considerably worse than the training ones, and simulation 8, MedCost, in which the difference of weights is 0.5, the results are consistent, and the both the training and test samples have similar results.

Table 3: Average Results of the Individual Simulation (part: II)

| FT 1 – FT 4 |
| FT 5 – FT 8 |
| FT 9 |
| FT 10 – FT 13 |
| FT 14 |

The individual test of the fraud cases simulated in Table 2 and Table 3, had, a positive training performance with low number of false detection and a high number true detection. The best situations are observed in the simulations 1 and 6, ValCost and UncMed respectively, both with a weight of 0.98.

In the simulation 1, all the GA used are extremely sensitive to changes in the costs of the claims, which explains the good results, almost total detection, obtained with the training sample. This situation is also observed in simulation 6, where is tested fraud type 6, UncMed, that mainly influences the medicine used. In this cases, the three GA consider and compare the high use of a specific medicine, which enabled a good training performance, above 0.9 of weight.

The worse simulated case scenario is the simulation 2, NumClaim, with just 0.46 of weight. The number of claims of a specific patient is neither detected by the claims GA, neither by the provider GA. The performance of the second GA, dedicated exclusively to the patient perspective, as an important influence in the detection of this fraud case, considering that the number of claims variable is directly evaluated by the genes of the chromosomes in this GA.

With the test sample the scenario does not change. The best situation is observed as expected in simulation 6, UncMed, with a weight of 0.91. And the worse situation in simulation 2 and simulation 8, NumClaim and MedCost.

The relation between the performance of the criteria’s, developed in the GA in the training sample and tested in the test sample, is observed with the delta weigh.

Excluding simulation 1, ValCost, in which the test results are considerably worse than the training ones, and simulation 8, MedCost, in which the difference of weights is 0.5, the results are consistent, and the both the training and test samples have similar results.

2) Group Simulation Part I

These group simulations were developed with a training population of 765 with 15 fraudulent claims, and a test population of 265 with 15 fraudulent claims.

Table 4: Average Results of the first population in Group Simulation

| FT 1 – FT 4 |
| FT 5 – FT 8 |
| FT 9 |
| FT 10 – FT 13 |
| FT 14 |

The individual test of the fraud cases simulated in Table 2 and Table 3, had, a positive training performance with low number of false detection and a high number true detection. The best situations are observed in the simulations 1 and 6, ValCost and UncMed respectively, both with a weight of 0.98.

In the simulation 1, all the GA used are extremely sensitive to changes in the costs of the claims, which explains the good results, almost total detection, obtained with the training sample. This situation is also observed in simulation 6, where is tested fraud type 6, UncMed, that mainly influences the medicine used. In this cases, the three GA consider and compare the high use of a specific medicine, which enabled a good training performance, above 0.9 of weight.

The worse simulated case scenario is the simulation 2, NumClaim, with just 0.46 of weight. The number of claims of a specific patient is neither detected by the claims GA, neither by the provider GA. The performance of the second GA, dedicated exclusively to the patient perspective, as an important influence in the detection of this fraud case, considering that the number of claims variable is directly evaluated by the genes of the chromosomes in this GA.

With the test sample the scenario does not change. The best situation is observed as expected in simulation 6, UncMed, with a weight of 0.91. And the worse situation in simulation 2 and simulation 8, NumClaim and MedCost.

The relation between the performance of the criteria’s, developed in the GA in the training sample and tested in the test sample, is observed with the delta weigh.

Excluding simulation 1, ValCost, in which the test results are considerably worse than the training ones, and simulation 8, MedCost, in which the difference of weights is 0.5, the results are consistent, and the both the training and test samples have similar results.

2) Group Simulation Part I

These group simulations were developed with a training population of 765 with 15 fraudulent claims, and a test population of 265 with 15 fraudulent claims.
Table 5: Average Result of the second Population in Group Simulation (part: I)

<table>
<thead>
<tr>
<th></th>
<th>Simu. 15</th>
<th>Simu. 16</th>
<th>Simu. 17</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FT 2 - 4 - 7</td>
<td>FT 1 - 3 - 9</td>
<td>FT 1 - 3 - 4 - 5</td>
</tr>
<tr>
<td>Avg</td>
<td>0.64</td>
<td>0.55</td>
<td>0.99</td>
</tr>
<tr>
<td>TP</td>
<td>0.13</td>
<td>0.08</td>
<td>0.07</td>
</tr>
<tr>
<td>FP</td>
<td>0.51</td>
<td>0.47</td>
<td>0.92</td>
</tr>
<tr>
<td>W</td>
<td>0.14</td>
<td>0.29</td>
<td>0.93</td>
</tr>
<tr>
<td>ΔW</td>
<td>0.05</td>
<td>0.16</td>
<td>0.09</td>
</tr>
<tr>
<td></td>
<td>0.08</td>
<td>0.13</td>
<td>0.84</td>
</tr>
</tbody>
</table>

Table 6: Average Result of the second Population in Group Simulation (part: II)

<table>
<thead>
<tr>
<th></th>
<th>Simu. 18</th>
<th>Simu. 19</th>
<th>Simu. 20</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FT 1 - 3 - 4 - 6 - 7 - 9</td>
<td>FT 2 - 4 - 6 - 8 - 9</td>
<td>FT 1 - 3 - 4 - 6 - 7</td>
</tr>
<tr>
<td>Avg</td>
<td>0.65</td>
<td>0.53</td>
<td>0.67</td>
</tr>
<tr>
<td>TP</td>
<td>0.20</td>
<td>0.13</td>
<td>0.04</td>
</tr>
<tr>
<td>FP</td>
<td>0.45</td>
<td>0.40</td>
<td>0.64</td>
</tr>
<tr>
<td>W</td>
<td>0.31</td>
<td>0.21</td>
<td>0.43</td>
</tr>
<tr>
<td>ΔW</td>
<td>0.05</td>
<td>0.10</td>
<td>0.37</td>
</tr>
<tr>
<td></td>
<td>0.38</td>
<td>0.29</td>
<td>0.27</td>
</tr>
</tbody>
</table>

Observe Table 5 and Table 6, the conclusion may be taken from two different perspectives. First comparing the results with the previous section, i.e. observing the effect of a population ten times larger. And a second, observe the effects of a larger number of fraud types. The true positive rate is considerably low in all the situations, excluding in the simulation 17. In the cases of the false positive it remained low in both the training and in the test sample in all the simulations.

Delta weight is high, excluding in simulation 17. This situation principally indicates that the principal factor, for a good detection of fraud, are the fraud type and not the number of fraud examples, neither the different types of fraud in the sample. In simulations 18 and 20 for example, the results are worse in simulation 18 just because of fraud type 9 MedDist.

C. Comparison of Alternative Algorithm Performances

Was possible to compare this algorithm performance with three different strategies present in papers [2], [7] and [9].

Using the method present in [7], Genetic Algorithm and K-Nearest Neighbor Method. The best result is described with 82.00% of success detection of fraud in the training set and 82.21% in the test set. In the worse cases the training set had 80.38% of success detection of fraud and the test set 76.69%.

Making the comparison with the algorithm developed in this thesis, in certain simulation was achieve 100% of true positive rate in both training and test sample. The worse with 73.00% of true positive rate in training sample and 20.00% in the test sample. In one hand, the results obtained in [7] are much consistent, i.e., in all the simulation the relation between the training and test in less than 5%. In the other hand, in some situation the GA developed in this thesis achieves better results, 100% of true positive rate, in both training and test sample.

In the paper [2], the results are present according to the accuracy of the results. The authors used three different strategies, Logistic Regression with 92.18%, Neural Networks with 95.73% and Classification Trees with 99.37%, of accuracy. In the GA developed in this thesis, the first set, between simulation 1 and simulation 9, the average accuracy is 92.1% in training and 86.8% in test. In the second set, between simulation 10 and simulation 14 the average accuracy is 89.6% in training and 89.1% in test. In the third set, between simulation 15 and simulation 20, the average accuracy was 88.9% in training and 87.7% in test.

The paper [2] uses a Multilayer Perceptron Neural Network method. In the training sets the results indicated a true positive rate of 0.734 and a false positive rate of 0.069 and an accuracy of 96.07%. The results of the accuracy in the simulation of this GA are average values of all the experiences, GA Claims, GA Patient, GA Provider and the four intersections. Naturally the average accuracy is lower than in the two papers, [2] and [9], presented above, because in some scenarios the success of detection is considerable low.

Observing the three comparisons is clear that this GA is at the same quality level of the purpose in the papers. Accuracy near the 90% indicates that this GA as a good performance in the detection of fraudulent situations. The complete comparison between the algorithm, i.e., which one is the better, is not possible to be done, because the databases are not the same, and principally the fields used in each database are, probably, different.

V. CONCLUSIONS AND FUTURE WORK

A. Principal Points of Conclusion

The simulated results expose the continuous evolution of the detections rate. For the first scenario, the individual simulation of fraud cases, the average value of detection is satisfactory, depending on the fraud type, exist both, the ideal detection situation (only all the fraudulent situations are considered as fraudulent) and the good scenario, where the majority of the signal detection are fraudulent. The important results in this first scenario is that at least one of the experiences describes a perfect situation. This subject suggest that if the fraud case to detect is correctly specified and the appropriated method of detection elected, i.e. chosen between the three GA or the respective intersection of the results, it’s possible to detect all the fraudulent situations, without false alarms.
For the second scenario, small group simulation with the first population of one thousand claims, the results are worse than in the first. Both the average and best experience results are, yet, satisfactory. However, there is a decrease of the true positive rate. Although in training the sample present a good performance, in test, the criteria of detection don’t adapt so well. Is important to emphasize, once more, that is the true positive rate that decreases, making this good results to reduce the search in large data bases.

The third scenario, large group simulation with the second population of ten thousand claims. The deterioration of the average results is evident, once again true positive rate decreases and the difference between training and test is considerable. However, the initial conclusion persists: If the fraud cases to detect are correctly specified and the appropriated method of detection elected, in specific fraud types the objective of detect all and only the fraudulent situation is achieved.

The number of different fraud scenarios influences the quality of the results. Samples with only one fraud type have ideal fraud detection situations. With the increase of the number of fraud types the quality of detection decreases. Until three different fraud types in the same populations the results are acceptable, with high true positive rate and a small false positive. The situation gets worse with the increase of the fraud types, tested until five different fraud cases.

Two of the most important points of this algorithm is first the use of three GA that process the same data base in different perspectives, claims, patient and provider. And second the intersection of the results of the GA. In general, is possible to understand that this were successful approaches to the problem. The specific genetic algorithms were capable of better detection in some types of frauds than other, which is normal considering the that the information observed from the perspective of the entities involved have different meanings. The intersection, were as well, successful approaches, principally in the simulations with a high number of different types of fraud. This experiments were able to reduce drastically the number of false positive detections.

B. Future Work

Considering the big disadvantages of this method, being the generator of data, the most important change is the simulation of the performance of the algorithm with real healthcare data. With the access to official information is possible in one hand to test the developed work and attribute veracity to the algorithm. In other hand, is possible to introduce another variable to the algorithm, improving the quality of the results.

Besides the efforts, the fraud types developed to test the algorithm is not perfect. It’s important to improve the already existing scenarios and create others taking example from healthcare reports of cases detected recently. The aggregation of fraud may also be integrated in the algorithm, i.e., the designed algorithm may be reorganizing in order to identify naturally the type of fraud and explores similarities in the database.

In this method were developed three different GA, would be interesting to understand which is more efficient to detect each type of fraud and the aggregation of the different types of fraud. As well as which of the intersection of the three GA is more capable of detect certain types of fraud.

VI. References