Applied Soft Computing xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

An interactive machine-learning-based electronic fraud and abuse detection system in healthcare insurance

Ilker Kose[a,*], Mehmet Gokturk[b,1], Kemal Kilic[c,2]

[a] Istanbul Medipol University, 34810 Istanbul, Turkey
[b] Faculty of Computer Engineering, Gebze Technical University, 41400 Kocaeli, Turkey
[c] Faculty of Engineering and Natural Sciences, Sabanci University, 34956 Istanbul, Turkey

ARTICLE INFO

Article history:
Received 22 June 2013
Received in revised form 26 April 2015
Accepted 5 July 2015
Available online xxx

Keywords:
Decision support systems
Fraud detection
Health insurance
Interactive machine learning
Machine intelligence

ABSTRACT

Detecting fraudulent and abusive cases in healthcare is one of the most challenging problems for data mining studies. However, most of the existing studies have a shortage of real data for analysis and focus on a very limited version of the problem by covering only a specific actor, healthcare service, or disease. The purpose of this study is to implement and evaluate a novel framework to detect fraudulent and abusive cases independently from the actors and commodities involved in the claims and an extensible structure to introduce new fraud and abuse types. Interactive machine learning that allows incorporating expert knowledge in an unsupervised setting is utilized to detect fraud and abusive cases in healthcare. In order to increase the accuracy of the framework, several well-known methods are utilized, such as the pairwise comparison method of analytic hierarchical processing (AHP) for weighting the actors and attributes, expectation maximization (EM) for clustering similar actors, two-stage data warehousing for proactive risk calculations, visualization tools for effective analyzing, and *z*-score and standardization in order to calculate the risks. The experts are involved in all phases of the study and produce six different abnormal behavior types using storyboards. The proposed framework is evaluated with real-life data for six different abnormal behavior types for prescriptions by covering all relevant actors and commodities. The Area Under the Curve (AUC) values are presented for each experiment. Moreover, a cost-saving model is also presented. The developed framework, i.e., the eFAD suite, is actor- and commodity-independent, configurable (i.e., easily adaptable in the dynamic environment of fraud and abusive behaviors), and effectively handles the *fragmented* nature of abnormal behaviors. The proposed framework combines both proactive and retrospective analysis with an enhanced visualization tool that significantly reduces the time requirements for the fact-finding process after the eFAD detects risky claims. This system is utilized by a company to produce monthly reports that include abnormal behaviors to be evaluated by the insurance company.

1. Introduction

In traditional machine learning applications, the domain experts (who are usually the *end users* as well) usually participate in the modeling stage at two points [1,2]. First, the domain expert educates the knowledge engineers who develop the decision support tool so that the tacit knowledge (e.g., the objectives of the machine learning application, the relevant features that should be utilized in the analysis, natural groupings that are already known by the experts, and how to address missing data, etc.) can also be included in the development process. Secondly, especially for applications where the readily available data lack the required labels for machine-learning-based classification algorithms (i.e., unsupervised learning), the domain experts are asked to label the training data. Both of these contributions are commonly used and have proven to be very successful in various real-life applications.

However, in certain domains such as defense, healthcare, biosciences, etc., the experts are highly motivated and want to interact with the data and tools; hence, they are not satisfied by the above-mentioned role [1,3]. Therefore, they are usually reluctant to use traditional machine learning systems in their analysis unless they engage in the development process and are able to customize it

[*] Corresponding author. Tel.: +90 216 681 53 056; fax: +90 212 531 75 55; mobile: +90 532 645 24 34.
E-mail addresses: ikose@medipol.edu.tr (I. Kose), gokturk@gyte.edu.tr (M. Gokturk), kkilic@sabanciuniv.edu (K. Kilic).
[1] Tel.: +90 262 605 22 09; mobile: +90 533 425 83 95.
[2] Tel.: +90 216 483 95 96; mobile: +90 505 211 82 99.

http://dx.doi.org/10.1016/j.asoc.2015.07.018
1568-4946/© 2015 Published by Elsevier B.V.

based on their expertise. Furthermore, although computers are able to analyze vast amounts of data, human intelligence might be still preferable for analyzing smaller data sets in more detail [1,4,5].

As a result, the concept of interactive machine learning (IML) has emerged and attracted the attention of the research community and practices due to its ability to incorporate the domain experts directly into the model building process by providing various human–computer interaction tools and data visualization and analysis techniques as part of the developed applications [6–8]. Using IML, users can train a learning algorithm by choosing parameters, then evaluate and compare models, selecting those that are most appropriate for their goals.

Unlike the traditional machine learning approach, in which the humans (users) and machine work independently on different tasks, in IML, they work on the same task during the training stage [1,5]. That is, the human component utilizes the computational power of the machines to learn from the relationships hidden in data (using data visualization and analysis techniques) and in return directs the training process. This interaction makes IML particularly valuable whenever the hypotheses and objectives of machine learning are subject to change. The users can interact with the execution of the tool and modify it within the deployed environment.

This current research is motivated by the demand for a reliable and usable tool for fraud detection in the healthcare insurance business. In this study, an IML-based decision support tool that couples claim management systems, which will be utilized to detect fraud and abuse, is developed. The decision support tool uses the transactional data in order to identify suspicious cases (by assigning a value-function-based risk measure to each transaction) and provides a visual environment to aid users in the determination of whether the transaction is actual fraud or abuse.

Although there are some studies on anomaly detection time-series data and event sequence data [9] note that it is not viable to decide whether a *single transaction* is fraudulent or abusive in healthcare claims other than the trivial cases (such as prescribing a postpartum drug to a male patient, etc.) [10]. Usually, fraud and abuse in healthcare claims can only be detected if the earlier transactions by the same *actors* are also taken into consideration during the analysis. On top of the fruitless process of trying to label the transactions as fraudulent or abusive, the abundant number of transactions also limits the applicability of more elaborate classical machine learning techniques as the engine of the developed decision support tool. Furthermore, the fraudulent and abusive behavior evolves over time [9,11]. That is to say, the *actors* of the system are intelligent and adapt to the policing of insurance claims by changing their tactics. Therefore, rather than classical machine learning techniques, an IML-based decision support tool is extremely well suited for detecting fraud and abuse in healthcare insurance.

The developed framework imitates the process that experts use to determine suspicious cases, which is usually based on a *bottom-up* analysis of each actor and his or her relationship with healthcare associated commodities. Next, a *top-down* approach is utilized to automate the experts' method to identify the relevant evidence. The developed decision support tool significantly decreases the necessity of manual analysis by highlighting only the most suspicious cases and eliminating those that are unlikely to be critical. Furthermore, the data visualization tool enables the user to investigate each case effectively and thereby learn more about fraud and abusive behaviors, informing modifications of the risk assessment and evaluation engine.

The rest of the study is organized as follows. In Section 2, we present the problem statement and the ecosystem of the fraud and abuse detection problem. This section is critical for readers who are not familiar with the healthcare insurance domain. Furthermore, it will also clarify why a single, perfectly normal-seeming transaction might in fact be part of a greater fraud. Section 3 is devoted to the review of the related literature. Note that the existing literature, which is subtle but growing, deviates from our research in that it focuses only on the fraudulent and abusive behavior of a particular *actor* and not the transactions, which incorporate multiple actors and commodities. The developed framework is introduced in detail in Section 4. In order to measure the performance of the framework, an experimental analysis based on real data is presented in Section 5. The study ends with some concluding remarks and future research plans.

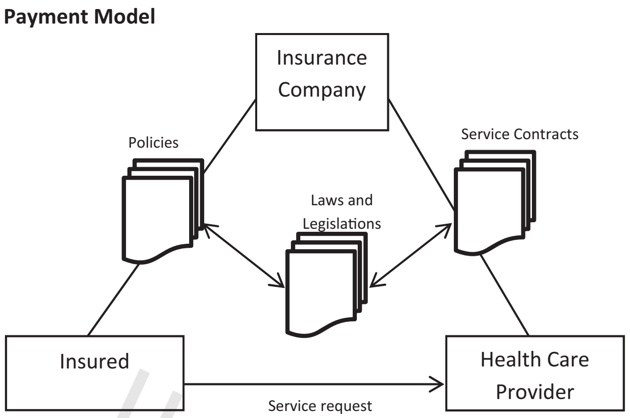


Fig. 1. The health insurance payment model.

2. The problem definition

The steady increase in life expectancy due to advancements in the health sciences, better standards of living, and increasing awareness of healthy lifestyles significantly enlarged the magnitude and share of healthcare in the global economy. For example, the Center for Medicare and Medicaid Services (CMS) reported that the total healthcare spending for the United States of America in 2010 was 2.6 trillion USD, which is 17.9% of the nation's gross domestic product (GDP). For comparison, in 2000, the US's total healthcare spending was 1.4 trillion USD (nearly half of its spending in 2010), and its share of the nation's GDP was 13.8% [12]. Improvements in medical technologies have also lead to major process and product diversification in healthcare services. Increase in healthcare expenditures, diversified healthcare services, and the rising expectations of the insured have cultivated the need for a systematic approach to manage the entire process. As a result, *healthcare insurance payment management systems* (for public and private insurance) have been developed to maintain the competitiveness of insurance companies and organizations in the marketplace.

Any healthcare insurance payment management system includes *the payment model* (e.g., fees for services, fees for capita) and *the claim management system*. The specifications regarding the payment, which are set by exogenous decision-makers, i.e., regulatory bodies, such as the government and the insurance authorities, primarily determine *the payment model*. Fig. 1 depicts the payment model which governs the payment among the three major stakeholders of the healthcare insurance systems, namely the insured, the insurance company and the healthcare providers. In addition, the *claim management system* component monitors the policy and contract terms (which are represented by the *payment rules*) among the stakeholders in order to provide proper healthcare service to the insured at the prices specified in the contract.

Generally, in practice, the management of the payment rules is considered as a relatively easy task and handled mostly by a

]

rule-based system in practice. On the other hand, the data accumulated by the claim management systems also provide the opportunity for more advanced analysis, such as fraud detection and management. However, detecting fraudulent and abusive behaviors is not an easy task and requires more intelligent solution approaches that are *intuitive* (for the users) [13] and heuristic-based due to the highly nonlinear and complex nature of the problem. Although various electronic *claim management systems* have been developed and successfully used to manage payment rules in the marketplace (e.g., ECP Drug of CompuGroup Medical) [14], no analogous product can manage fraud detection in the marketplace, other than the few isolated attempts reported to date (which will be discussed in more detail in the literature review).

Many health insurance payment management systems primarily rely on human experts to review insurance claims and identify suspicious cases manually [15]. Because manual review of suspicious claims is quite an expensive way to detect fraudulent and abusive behaviors and has a highly questionable performance [11,16], insurance companies are attempting to develop an electronic fraud and abuse detection system that utilizes actual data gathered via *claim management systems.* Thus, researchers have been increasingly interested in fraud and abuse detection research in recent years especially that use techniques based on data-mining methodologies.

The National Health Care Anti-Fraud Association (NHCAA) defines healthcare fraud and abuse, respectively, as follows [17]:

> "*Health care fraud is an intentional deception or misrepresentation that the individual or entity makes knowing that the misrepresentation could result in some unauthorized benefit to the individual, or the entity or to some other party*."

> "*Health care abuse is the provider practices that are inconsistent with sound fiscal, business and medical practices, and result with unnecessary cost, or in reimbursement of service that are not medically necessary or that fail to meet professionally recognize standards for health care*".

U.S. Federal Bureau of Investigation [18] estimates that between 3% and 10% of claims in both public and private healthcare expenditures are fraudulent. Due to the considerable increase in healthcare spending experienced by most countries, efficient handling of fraud and abuse detection is crucial for insurance companies, as they struggle to decrease healthcare costs while ensuring that the insured are adequately satisfied with their coverage.

In order to understand the fraud detection problem within the context of healthcare insurance, it is vital to understand the healthcare insurance payment process and claim management tools utilized during this process, i.e., electronic claim processing (ECP).

2.1. The role of electronic claim processing (ECP) systems

There are four key entities in a claim management system: the insurance company, the beneficiaries (i.e., the insured), the service providers, and the claims [19]. The claim management system validates the claim, i.e., checks whether it is appropriate to make the payment based on the terms of the contracts between the stakeholders. These conditions cover a variety of factors, such as the content, amount, price, and whether the individual is covered for the particular sickness. To monitor the claim process, the health care provider should provide necessary information via the claim submission. This set of data, which is referred to as the *electronic claim data set* hereinafter, includes factors, such as the identity of the insured, content, amount, and price.

Most insurance companies use basic ECP systems, which generally use rule engines in claim management. These rule engines include rules that are primarily based on simple criteria, such as gender, age, dependent services, mutually exclusive services and services that are beyond the scope of the policy. An ECP system deterministically evaluates the particular claim itself and does not refer to earlier claims by the same beneficiary and/or the health provider that is involved in the particular claim. Rare exceptions occur for certain basic quantity controls, such as "allowed *x time(s)* in a year", which are still rule-based.

There are two categories of rules that ECP systems handle: medical rules and insurance rules. The medical rules indicate whether the provided health service is appropriate for the declared age, sex, and similar attributes of the insured, whereas the insurance rules are the common and specific terms specified by the contract (policy) of the insured. Since the goal with the ECP systems is not to replace provision experts but to support their function [20], they cannot handle all of the medical and insurance rules completely. First, the medical rules are very difficult to model for all possible cases. Second, the electronic claim data set includes very limited data due to the practices and unwillingness of the healthcare providers. Third, the nature of medical science, particularly the wide variety of alternative approaches in the treatment process, requires a vast amount of knowledge and flexibility during the validation stage. Lastly, the change management of the available rules can be dramatically sophisticated for large numbers of rules. Therefore, nearly all health insurance companies establish a claim management office, wherein a team of medical experts struggles to manage those cases that are not handled by the ECP system. The general usage of ECP systems and their relationship with claim management offices is depicted in Fig. 2.

Even though effective management of the payment rules faces the above-mentioned challenges, the ECP systems still primarily succeed using a rule engine. As a result, the ECP systems accumulate a huge set of data, which can be mined for various other reasons, such as fraud and abuse detection. However, this task is even more challenging due to the peculiarity of fraudulent and abusive behaviors. In particular, the fragmented nature of fraud and abuse behaviors requires more intelligent approaches and extensive computational resources for detection.

2.2. The fragmented nature of fraud and abuse behaviors

Fraud and abuse behaviors are generally fragmented into a series of claims and can only be identified after defragmentation [10,21]. That is, it is very hard to decide whether a *single transaction* is fraudulent or abusive; instead, the earlier transactions by the same *actors* must also be taken into consideration in the analysis. Consider a fraud scenario in which a physician and a pharmacist fix a deal with a drug company representative so that the physician prescribes drugs from that particular company for various patients with or without their knowledge, i.e., creates *fake prescriptions*. It is quite possible that every prescription can be considered to be valid with respect to the payment and medical rules by the ECP systems. However, one can compare the activities of the actors (i.e., physicians, pharmacists, and beneficiaries) with other similar actors or with self-historical transactions with regard to the particular drug company and can determine the anomaly in such fake transactions. That is, a transaction that appears very normal can in fact be part of a set of fraudulent transactions within a period. Therefore, a thorough analysis is required for an effective fraudulent and abusive behavior detection process. Fig. 3 illustrates the fraudulent behavior of the fake prescription and its fragmented nature. Note that the gray prescriptions in the figure might be valid transactions individually but, when taken as a group, might reveal a plot among the actors.

The fragmented nature of fraud and abuse behaviors limits alternative solution approaches to the fraud and abuse detection problem and mostly excludes supervised learning-based machine-learning approaches. Note that the supervised learning methods

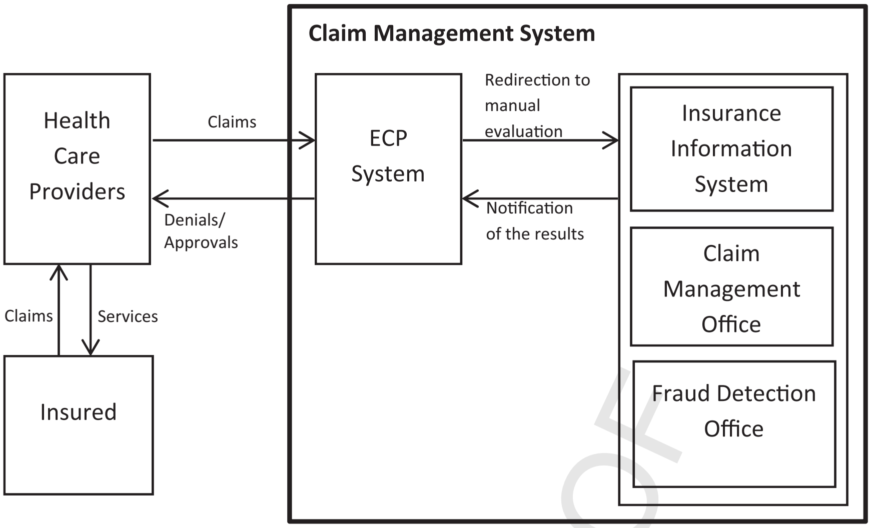


Fig. 2. The electronic claim management systems.

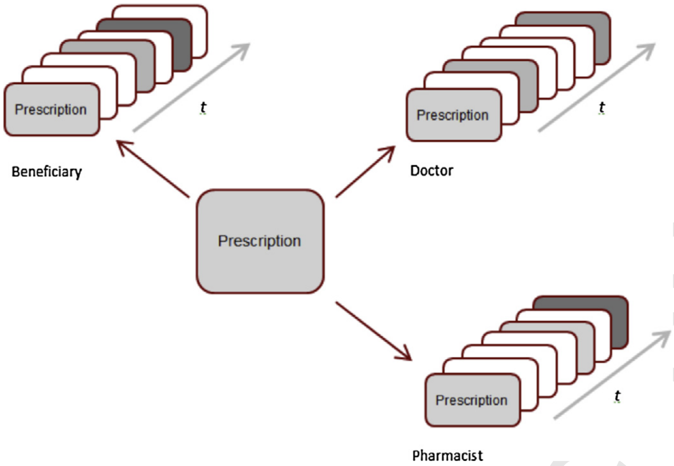


Fig. 3. The fragmented nature of fraud and abuse behaviors.

require each transaction to be labeled as suspicious or not suspicious (e.g., by domain experts). However, a *transaction should not* be labeled as fraudulent or abusive in the first place. For example, it is quite possible that for two transactions with precisely the same content in terms of the electronic claim data sets, one could be part of a *fraud* and the other *not*, depending on the *earlier* transactions of the actors involved in each transaction [10,22].

2.3. The ecosystem of the fraud and abuse detection problem

Before presenting the formal definition of the fraud and abuse detection problem, the following definitions must be presented:

Definition 1. An ***actor*** A_i^b is a uniquely represented participant involved in a claim transaction. An actor type (specified by the b index in A_i^b) can be an insured individual, physician, or institutional health provider, such as a hospital or pharmacy.

Definition 2. A ***commodity*** $\boldsymbol{C_j}$ is a uniquely represented treatment material, such as a medication or health service, that the insurance company is requested to pay for. The commodity can have tightly coupled attributes so that, they can be used instead of the commodities' brand name in order to represent the relationship between actors and commodities. For example the Anatomical Therapeutic Chemical (ATC) code, which is used to define the active ingredient of the drug, can be used instead of the brand name of the drug. This enables to analyze the relations of actors with commodities from different perspectives.

Definition 3. A ***medical claim*** $\boldsymbol{MC_k}$ is a set of triplets ($\{A_i^b\}$, $\{C_j\}$, t), where $\{A_i^b\}$ and $\{C_j\}$ identify the set of actors and commodities involved in the claim, respectively, and t represents the time, e.g., transaction.

The key *actor* type in the process is the patient, and the objective of the rest of the *actors* is to diagnose and treat the patient. For both of these purposes, the patient can visit several healthcare providers (several times each) during the whole process, and each visit that involves a payment results a *medical claim*. All of the payments are for *commodities* supplied to the patient by the healthcare providers. Therefore, *any* fraud and abusive behavior should be associated with those *commodities* to identify unauthorized benefits by one or more of the *actors* involved in the process. Note that as discussed earlier, identification of any *abnormal behavior* (e.g., fraud and abuse) requires an analysis of fragmented medical claims as a whole.

Definition 4. An ***abnormal behavior*** $\boldsymbol{F_l}$ is a fraudulent or abusive behavior of the actors that can be determined based on a set of fragmented medical claims $\boldsymbol{MC_k}$.

There are different types of abnormal behavior, depending on the actors and commodities involved. For example, the National Health Care Anti-Fraud Association (NHCAA) [23] specifies the following types of abnormal behavior (among others) for various actors:

The *service provider's* fraud and abuse, including

- Billing services that were never rendered.
- Performing more expensive services and procedures than necessary.
- Performing medically unnecessary services solely for the purpose of generating insurance payments.
- Misrepresenting non-covered treatments as medically necessary for the purpose of obtaining insurance payment.
- Falsification of patients' diagnosis and/or treatment histories.

The *insured's* fraud and abuse, including

- Misrepresenting an application to obtain a lower premium rate.
- Falsification of records of employment/eligibility.

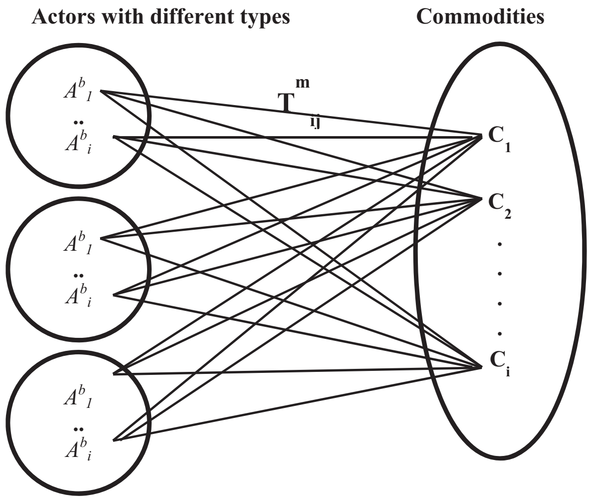


Fig. 4. Attributes representing the relation between actors and commodities.

- Falsification of medical claims.

The *insurance company's* fraud and abuse, including

- Falsification of reimbursements.
- Falsification of benefit/service statements.

Definition 5. The ***attributes*** $T^{m,b}_{ij}$ are metrics representing the scores of the relationship between actors A^b_i and commodities C_j. Note that different types of fraudulent and abuse behavior produces anomalies in various attributes as indicators of their existence; therefore, the index *m* represents the associated metrics as depicted in Fig. 4. The figure illustrates that the attribute $T^{m,b}_{ij}$ refers to the score of the *m*th metric that measures the relation between the *i*th entity of the *b*th actor type and the *j*th commodity.

Definition 6. The ***proactive approach*** is a way of detecting fraudulent and abusive cases on-line. Such systems are used to support the medical insurance experts who work in the provision departments to approve the claim or not in an acceptable time and show the risk degree of the claim and relevant justifications of the risk of the actors.

Definition 7. The ***reactive approach*** is a way of detecting fraudulent and abusive cases retrospectively. Such systems are used to support the damage experts who work in the reimbursement departments to approve the payment to the health institutions or not. Such systems are not required to process the risks on-line, but the reporting and analysis should be fast enough for the users. Such systems should also show the risk degree of the claim and relevant justifications of the risk. If there is a proactive system that can calculate the risk of any claim on-line, then the reactive system may use the risk engine of the proactive system to calculate the risks; but shows the risky claims in more comprehensive analysis module to provide the damage experts to make more detailed evaluation.

In summary, the fraudulent and abusive behavior detection problem involves the determination of whether a medical claim demonstrates a particular abnormal behavior given the earlier fragmented medical claim data set that involves all relevant actors and commodities by considering the dynamic nature of the problem.

3. Review of the relevant literature

Although, as mentioned above, the prevalence of fraudulent claims is estimated to be between 3% and 10% in both public and private healthcare expenditures [18], a limited number of studies have addressed fraud detection in healthcare insurance. There are various reasons for this shortage. First of all, fraud detection requires an extensive amount of data as well as data processing resources. However, electronic data has only become commonly available in the last decade, with the advent of telecommunication technologies. Secondly, the fraud detection process in the healthcare context is more complex than that in other contexts, such as credit card fraud or car insurance; thus, the motivation of investors to fund research is limited. Thirdly, even though a certain degree of research has been conducted on limited versions of the healthcare insurance fraud detection problem, the details of the approach are usually a business secret; therefore, it is difficult to report the results in a way that satisfies the curiosity of the scientific community without divulging too much information. However, we can confidently state that there are virtually no healthcare fraud detection products available as a viable solution for the complete problem in the marketplace.

Phua et al. [24] present a comprehensive literature review of fraud detection problems, covering 51 studies that incorporate data mining approaches for fraud detection problems. Among these 51 studies, only 14 are associated with fraud detection in the context of insurance, and only five are in the healthcare insurance domain. The same study states that researchers usually complain about a lack of data to be analyzed and a shortage of well-examined methods and techniques in the published literature. However, only seven of the 51 examined studies have actually been implemented (only two of these are in the insurance business and none are in healthcare insurance).

A more recent review focuses on data mining techniques in financial fraud detection only and covers 49 papers, suggesting increased interest by researchers [25]. Among these 49 papers, five are in the domain of healthcare insurance (three overlap with those covered by Phua et al. [24]).

The first comprehensive healthcare fraud detection literature review (published prior to the previous two reviews) is presented by Li et al. [26]. This review classifies studies in terms of the utilized feature selection and statistical modeling techniques, performance evaluation approaches, data sources, and data pre-processing. The review also provides a discussion of the limitations of the existing techniques and presents various challenges for future research. In this section, we will briefly discuss the studies that were covered in these reviews as well as those published later.

He et al. [27] target the fraud and abusive behavior of general practitioners (GPs) utilizing the Australian Health Insurance Commission (HIC) data. These authors use 28 features that experts had suggested to be relevant in the fraud detection process. The experts also specified whether existing data should be tagged as fraud or abusive behavior, facilitating a supervised learning approach. In this study, a multi-layer perceptron (MLP) neural network was used to classify the practice profiles of a sample of 1500 general practitioners. Next, the effects of using a self-organized map (SOM) in conjunction with MLP were examined. SOM was used to classify the network classes, after which the MLP was applied for two-class classification, yielding better results. The authors report that the agreement rates (i.e., accuracy) were 63.60%, 59.87%, and 88.40% for MLP, SOM, and SOM followed by MLP, respectively. Note that this study only addresses data regarding GPs and does not attempt to develop a fraud detection method for the transactions.

Williams [28] conducts another study based on the HIC database. As a solution for fraud detection in the health insurance sector, Williams proposes a data mining method referred to as *hot spots* (*HS*). The HS method is fundamentally based on the determination of interesting and analyzable nuggets (i.e., chunks). The methodology is proposed for data mining applications in a general context, and healthcare fraud detection is used as a case study.

Please cite this article in press as: I. Kose, et al., An interactive machine-learning-based electronic fraud and abuse detection system in healthcare insurance, Appl. Soft Comput. J. (2015), http://dx.doi.org/10.1016/j.asoc.2015.07.018

In the analysis, over 30 raw attributes (e.g., age, sex) and approximately 20 derived attributes (e.g., number of times a patient visited a doctor per year, number of different doctors visited) are utilized to detect fraud and abusive behavior on the part of patients. A three-step framework of unsupervised learning is developed. In the first step, the dataset is clustered with a multivariate *k*-means algorithm. The second step is rule induction, in which one or more rules are constructed for each cluster attained in the first step. Lastly, an interestingness score is calculated for each rule. In the healthcare fraud case, the average number of services and average total benefit paid to patients, among other factors, are used to calculate the interestingness score. In the paper, no experimental analysis in terms of model validation is presented, and the authors declare that, although early feedback indicates that HS does provide a useful expansion of the search space, they cannot claim the usefulness of the method. The proposed framework can also be generalized for other actors (e.g., physicians or pharmacies), having different relevant features. However, again, the study is limited to the *actors* and does not consider the transactions.

Yamanishi and Takeuchi [29] propose an online-unsupervised outlier detection methodology named SmartSifter. In their work, the HIC database is utilized as a case study for the SmartSifter application in healthcare fraud detection. SmartSifter addresses the problem from a statistical learning theory viewpoint. Each time a new data point is fed, SmartSifter evaluates how much the data point deviates from the expected value, which is calculated based on a probabilistic model learned from the existing dataset. The probability density of the categorical values is determined using a histogram, and the density of the continuous variables is determined based on a Gaussian mixture model. Even though only fraud and abuse behavior detection for the pathology providers is presented in the paper, the proposed methodology is also applicable for various other actors, such as patients and physicians (left as a future research topic). For the case of pathology providers, only seven features were utilized (five of which were proportions in five different pathology groups, i.e., microbiology, chemical, etc., and the sixth was the number of different patients) in the analysis. The proposed methodology can be used in both proactive and reactive settings. However, the risk of the actor rather than the transaction is addressed. Furthermore, the authors did not conduct a formal experimental analysis of the proposed methodology, merely providing anecdotal evidence for the validation purposes.

Major and Riedinger [30] propose a two-stage *electronic fraud detection* (*EFD*) system as a reactive tool to identify suspicious *healthcare providers*. The EFD system incorporates 27 features (referred to as behavior heuristics in the paper) determined by experts in five different categories: financial, medical, logical, temporal, and spatial. For each health provider, samples associated with each feature are collected, and the sample statistics, such as the mean and variance, are calculated. The authors assume normality and use the sample statistics to determine the information gain of the corresponding feature, which is a measure of the deviation of a particular health provider and its peers. Next, the system utilizes a Pareto frontier curve, which is based on the total dollars paid, and the information gain to identify the most suspicious healthcare providers. The Pareto curve analysis is conducted for each of the 27 features, and those providers that are on the frontier of at least four features (i.e., the *threshold for hot tips* = four) are referred to as hot tips. The hot tips are brought to the attention of the investigative consultants for field investigations. The validation of the proposed system was not presented because the results of field investigations were unavailable at the time of publication. However, a receiver operating characteristics (ROC) curve analysis is presented based on the threshold for *hot tips frontiers*, assuming that all of the candidates examined by the investigative consultants are fraudulent. Although the authors do not give the AUC value, we can easily calculate it from the given ROC values as 66.53%.

Ortega et al. [31] propose a supervised learning methodology based on MLP to determine fraud in the context of medical reports for employee sick leave. Thus, this study is distantly related to our study but is included for the sake of completeness. In their methodology, four different models are developed for each actor in the process: the beneficiary (i.e., the employee), the physician (who prepares the medical report), the employer, and the medical claim. In all, 125 different features associated with the four actors are used in their model. The fraudulent and abusive behaviors of each relevant actor are inferred based on the values of their features. The proposed methodology is applied in a case using two datasets acquired from Banmedica (Chile). In the datasets, each claim is tagged as accepted, rejected, or reduced by the experts. Only partial information is presented due to the terms of a confidentiality agreement with Banmedica. To validate the proposed model, ROC analysis is utilized. The authors suggest an optimal threshold value yielding a true positive rate of 73.4% and a false positive rate of 6.9%.

Yang and Hwang [15] utilize the concept of clinical pathways (i.e., care plans in which diagnosis and therapeutic intervention are performed by physicians, nurses, and other staff for a particular diagnosis or procedure) to determine fraud and abusive behavior on the part of healthcare providers. In their proposed methodology, a graph is formed considering the reference pathway data for a certain disease. Nodes on the graph demonstrate the processes that are parts of the clinical pathway, whereas arcs represent the precedence relationships among the processes. Next, all possible single, double, triple, etc., sub-graphs of this graph are determined and utilized as features in their methodology. However, the clinical pathway of an average disease yields tens of thousands of sub-graphs. Therefore, a filter-based feature selection approach is utilized as a part of the supervised learning process, where the experts specify the labels of the data. Lastly, the C4.5 algorithm is used for classification purposes. To validate the proposed methodology, the data set associated with pelvic inflammatory disease (PID) from the gynecology department of a Taiwanese hospital is used. The rate of correct detection of cases, including fraud or misapplication (i.e., the sensitivity) is 64%, whereas the rate of correct detection for cases that do not include fraud (i.e., specificity) is 67%. Again, the focus of their research is only distantly relevant for the purpose of our study.

Sokol et al. [32] use information visualization techniques to detect fraud and abusive behaviors. In their paper, special attention is given to the data preprocessing stage and data extraction, data transformation, and data auditing, which are presented in detail. Various examples of information visualization techniques are discussed in which field experts determine the analyzed features. The discussed techniques are appropriate for reactive analysis, and they do not provide any comparisons or validations of the results. They propose various machine-learning approaches, such as abuse and misuse profiling, normative profiling, and link analysis as future research topics.

One of the most recent studies on prescription fraud is that by Aral et al. [33]. This study proposes a novel approach to assess the fraudulent risk of prescriptions (i.e., transactional data) based on cross-feature analysis, which can be used as both a reactive and a proactive tool. In their proposed approach, five pairs (i.e., medicine–diagnosis, medicine–age, medicine–sex, medicine–medicine, diagnosis–cost) are chosen from six features (medicine name, price, prescription ID, age, sex, and diagnosis) based on correlation, and the corresponding incidence matrices are developed. These incidence matrices are used to calculate the risk matrices. The authors propose two different risk metrics, one for categorical features and one for ordinal features, yielding higher risk values with a decreased incidence rate. The authors use ROC curves and the area under curve (AUC) measures to validate the proposed techniques and report an AUC of 85.7%. To develop the ROC curve, a medical expert labels each prescription as fraudulent or not fraudulent, and the proposed algorithm's predictions are compared with these labels.

Another recent study proposed by Johnson and Nagarur [34] has a six-staged approach including provider profiling, demographic screening, claim amount screening, fraud risk quantification, risk threshold determination for fraud detection and comparison of risk values with risk thresholds. The prominent feature of this study is the usage of real data set including 878,691 claims of an insurance company. The study considers the doctors as the only actor type from four different specialties, such as otolaryngology, general practice, neurology, and ophthalmology. The accuracy model of the study is based on sensitivity, specificity and accuracy rates and measurements are held for all six stages and for four actor groups separately. Regarding the overall rates, the results vary from 83% to 88%, and the accuracy is %86 in average. At the end of the study, the authors compare the results of their study using ANOVA with the result of both unsupervised and semi-supervised neural networks. They claim that the mean accuracy rate of their study is significantly higher than both neural networks.

In summary, a general assessment of the above-mentioned studies, which are relevant to our research to varying degrees, is presented in Table 1. The deficiencies of the existing literature can be summarized as follows:

- Using only reactive methodologies is not a deficiency; however using both reactive and proactive approaches is an enhancement on a reactive only methodology. From this viewpoint, only two out of the eight studies claim some sort of proactive solution.
- The processing times, i.e., computational requirements are extremely high.
- Only one actor or process is considered in the analysis; however, in reality, fraud is often the joint work of multiple actors, such as the insured, the physician, the pharmacist, and none of them can be excluded from the analysis.
- No single model covers all types of claims, actors, and processes.
- Only one type of fraudulent and abuse behavior is taken into account in the studies.
- The relationship between actors and commodities has not been included in the analysis.
- Information on the results and the applications of the models in real-life cases is very limited.

As a result, taken individually, none of the previously mentioned studies would be adequate to detect fraudulent and abuse cases that involve multiple claims independently of the health providers, beneficiaries, and type of service. The absence of commercial fraud detection systems in the healthcare insurance domain also supports this observation.

4. Methods

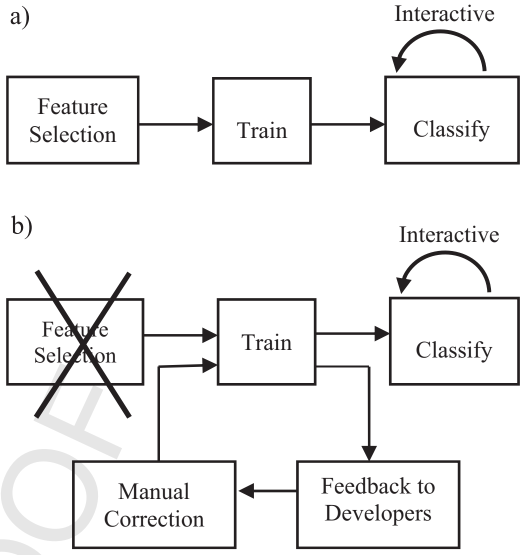


Fig. 5. (a) The classical machine learning process. (b) The interactive machine learning process [2].

The developed framework is a *novel approach* that attempts to overcome the shortcomings of the existing research. Firstly, the developed framework is *independent* of individual *actors* and *commodities.* That is, the fraud cases handled are not limited to the behavior of a single actor type (e.g., those perpetrated by physicians) or a set of commodities; rather, multiple actor types and commodities are considered simultaneously. Secondly, the way in which fraud and abuse behaviors are conducted will change over time in terms of the types of actors and commodities, i.e., the fraud ecosystem is *dynamic*. Therefore, a flexible framework based on the interactive machine learning approach is adopted which is capable of handling such changes. The IML method also enables experts and/or users to interact with the training process and contribute to the modeling with their expertise. Note that IML method exists in the literature and the novelty in our case is its application and adaption in the context of healthcare insurance. Thirdly, we used well-known methods within our framework, such as pairwise comparison [38] for weighting the actors and attributes, expectation maximization method [36] for clustering the similar actors, the *z*-scores of the attributes generated from the historical transactions for determining the risks of the claims, two-staged data warehousing for improving the performance to achieve proactive analysis and a dashboard-based technology (QlikView^{TM}) for visualizing the information. Fourthly, the framework enables the analysis of the *fragmented* claims as a whole. Recall that it is quite possible that for two transactions with precisely the same content, one could be part of a *fraud* and the other *not*, depending on the *earlier* transactions of the actors involved in each transaction. A fraud detection system that utilizes the earlier transactions of the actors has the potential to detect such cases as well. Lastly, the developed framework is a risk-score-based approach; however, unlike other risk score-based approaches that assign a risk score to the actors or commodities separately, the proposed risk scores are evaluated for *the actor–commodity relationship* as a whole. Therefore, even though an actor may be risky in terms of certain commodities, *only* those transactions that involve particular commodities and actors are labeled as fraudulent or not.

To address these issues, the electronic fraud and abuse detection (eFAD) framework, which is based on IML, is developed. Fig. 5 depicts the difference between the flow of the classical machine learning process and the flow of IML as presented by Fails and Olsen [2] in one of the pioneering studies for IML. The experts' (users') interactions with the tool during the training and classification stage will be explained next.

Before moving on to the discussion of the details of the process, the modules of the developed framework will be presented to enhance the reader's understanding. Fig. 6 illustrates the developed framework, which consists of four major components. The first component comprises the phase where the experts' knowledge regarding the objective and hypotheses of the machine learning process are incorporated. This is achieved via the storyboards generated by a team of insurance and medical experts. A two-stage data

Please cite this article in press as: I. Kose, et al., An interactive machine-learning-based electronic fraud and abuse detection system in healthcare insurance, Appl. Soft Comput. J. (2015), http://dx.doi.org/10.1016/j.asoc.2015.07.018

Table 1
Comparisons of previous studies.

Publisher	Type of reaction	Approach	Actors and processes	# of Attributes	Method	Architecture
He et al. (1997)	Reactive	Actor based	Doctor	28	MLP, SOM and MLP with SOM	Multi Layered Perceptron-(MLP) structured as 28-15-4 neurons and SOM in conjunction with MLP
Williams (1999)	Reactive	Actor based	Beneficiary	50	Hot spots	Anomaly detection framework with three steps including K-means clustering
Yamanishi et al. (2004)	Proactive (but with very high response time)	Actor based	Pathology service provider	41 (but four of them were used)	Statistical learning	The probability of the processes are calculated by a special method named as SmartSifter
Major and Riedinger (2002)	Reactive	Actor based	Hospital	27	Machine learning	A rule based model including operational and development phases
Ortega et al. (2006)	Proactive	Process based	Health report	125	2-Layered neural network classifier	Four sub-models of actors and claims generated from relevant DBs are used to feed the committee of multilayered feed-forward NN
Yang and Hwang (2006)	Reactive	Rule based (wrt clinical guidelines)	PID disease in Gynecology clinics	NA	C4.5 classifier	Classification of processes according to the graphs representing the clinical guidelines
Sokol et al. (2001)	Reactive	Process based	NA	NA	Visualization	A model including three phases, data cleaning, analysis and monitoring/visualization
Aral et al. (2012)	Proactive	Process Based	NA	6	Supervised learning	Uses normalized risk function based on distance calculations of the cross-features
Johnson et al. (2015)	Reactive	Actor based	Otolaryngology, general practice, neurology, and ophthalmology	NA	Density and distance analysis	Uses a six-staged framework based on Relative density ratio and distance calculations

warehouse is the second component, which is constructed based on the actors, commodities, and attributes extracted from the storyboards generated in the first component. The third component is the risk assignment engine, which evaluates the risk scores of the actors and the claims. The fourth component of the framework is the visualization tool, which produces various analyses based on the input values of attributes or the resulting risk scores of the claims and/or actors to the user. This visualization tool also enables the users to interact with the training of the decision support tool by changing parameters and introducing new relations as risk indicators for the transactions.

4.1. The story board: incorporating the experts' knowledge

In the current practice (i.e., without the use of a developed decision support tool), a team of medical and insurance experts attempts to determine abnormal behaviors merely using a reporting tool. The experts attempt to find an indicator of abnormal behavior based on various statistics (e.g., ratios, indicators, and trend lines) obtained from the existing database. As soon as the expert notices irregular behavior, he/she focuses on the actors and/or commodity involved in the particular irregularity. These actors and/or commodities are evaluated in terms of the medical and insurance rules and identified as *"risky"* or not based on these analyses. Depending on the particular situation, different managerial and legislative actions are taken next.

The developed framework mimics an expert's approach to assessing the risk of a medical claim. For this purpose, a storyboard is constructed, allowing for a team of medical and insurance experts to describe how certain known abnormal behaviors are realized. The storyboard specifies the actors, commodities, and attributes involved, representing the actor–commodity relations that yield abnormal values because of a particular abnormal behavior category. Based on the storyboard, the relevant actors, commodities, and attributes are extracted for each fraud type. In the framework, three modes of attributes are utilized, *periodical, differential*, and *cumulative*, due to different abnormal behavior types. Whereas the periodical attributes are monthly values and the cumulative attributes are yearly values, differential attributes represent the rate of change of a periodical attribute with respect to the previous month.

A storyboard study is presented as an example as follows. We first create a scenario of fraudulent or abusive behavior, which is expected to be representative of actual fraud. Assume that we would like to detect fraud perpetrated by a pharmacist, where the pharmacist knows the insurance card number of some insured persons and the registration number of some doctors. Next, he/she arranges fake prescriptions for the insured on behalf of the doctors. In addition, assume that the pharmacist is primarily arranging fake prescriptions (using the doctors' and patients' ID numbers) so that they include analgesic pills which are very cheap and usual in order to make his/her behavior seems as ordinary claims. In this case, our actors are the pharmacist, doctor, and insured, even though the doctor and insured are not intentionally participating in this behavior. The commodity is the drug, and we can calculate the relationship between the actors and drugs with respect to the Anatomical Therapeutic Chemical (ATC) codes in order to distinguish the fraud conducted with analgesics from other cases. To illustrate how the attributes are selected, certain attributes of the pharmacist representing the relationship between him/her and the commodities (with respect to ATC codes) are listed below:

- *The ratio of the number of prescriptions to the distinct number of insured persons*. Because the pharmacist is using the card

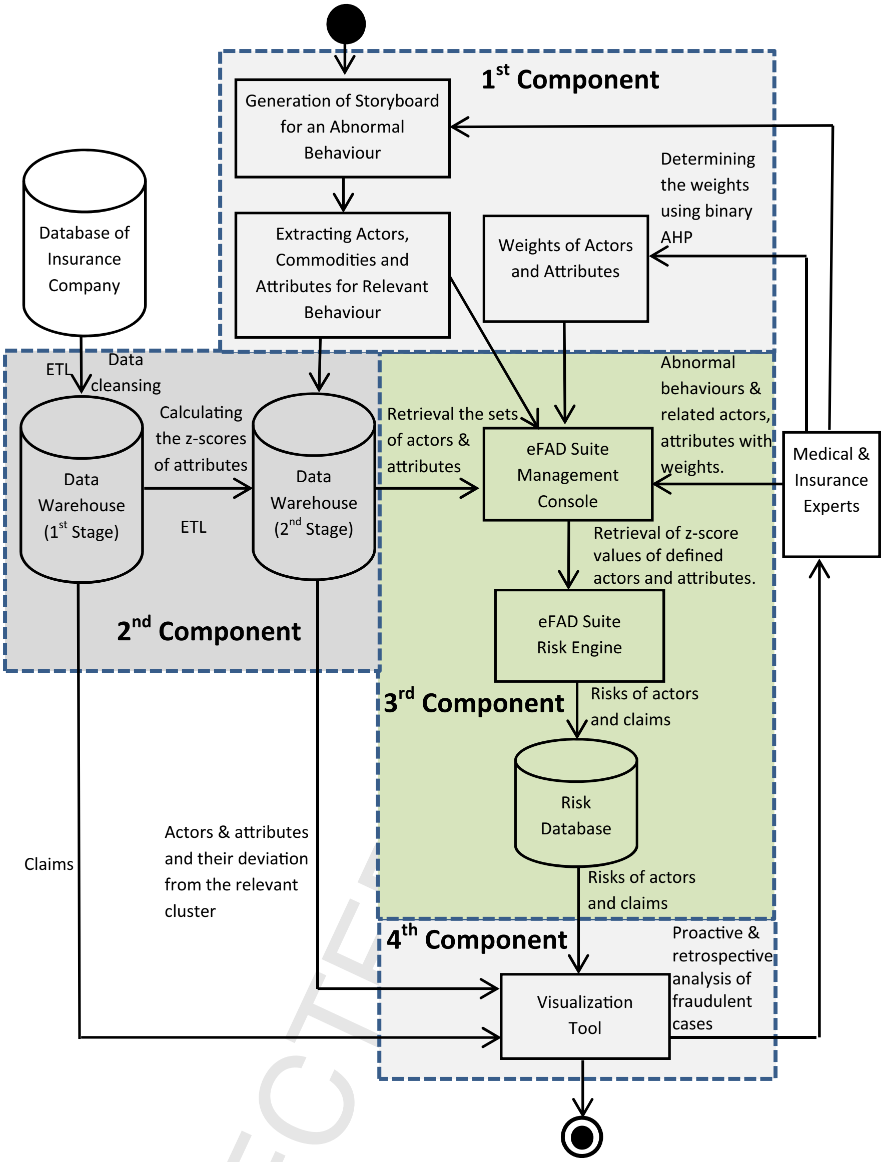

1st Component
Generation of Storyboard for an Abnormal Behaviour
Determining the weights using binary AHP
Database of Insurance Company
Extracting Actors, Commodities and Attributes for Relevant Behaviour
Weights of Actors and Attributes
ETL
Data cleansing
Abnormal behaviours & related actors, attributes with weights.
Calculating the z-scores of attributes
Retrieval the sets of actors & attributes
Data Warehouse (1st Stage)
Data Warehouse (2nd Stage)
ETL
eFAD Suite Management Console
Medical & Insurance Experts
Retrieval of z-score values of defined actors and attributes.
2nd Component
eFAD Suite Risk Engine
3rd Component
Risks of actors and claims
Risk Database
Actors & attributes and their deviation from the relevant cluster
Claims
Risks of actors and claims
4th Component
Proactive & retrospective analysis of fraudulent cases
Visualization Tool

Fig. 6. Structure and components of eFAD suite.

numbers of a small number of insured persons to arrange fake prescriptions, his/her insured patients, i.e., *distinct insured number*, will not increase linearly with his/her prescription number. This causes an abnormal increase in this ratio with respect to the ratios of other similar pharmacies in the same term. Here, we can choose a *periodical* attribute mode.

- *The ratio of the number of prescriptions to the distinct number of doctors*. Because the pharmacist is using a few doctors' registration information to arrange the fake prescriptions, his/her distinct number of doctors will not increase with his prescription number. This causes an abnormal increase in this ratio with respect to the ratios of other similar pharmacies in the same term. Here, we can choose a *periodical* attribute mode.
- *Total number of prescriptions*: If the pharmacist arranges those prescriptions within a short time, his/her total prescription number (including the relevant ATC codes) will increase unexpectedly relative to other similar pharmacies in the same term. We therefore should use the *differential* attribute mode to detect large changes in his/her prescription number within the term that the abnormal behavior occurred.

The attributes of other actors are extracted in the same manner. Notice that, when the storyboards of each fraud type are completed, it is clear that each fraud type may differ in both the type and number of actors, commodities, attributes, and weights. The second point that we should to notice is how to determine which values are normal and which others are abnormal. The answer of this question is stated in Section 4.2 as we use *z*-scores among the actor clusters.

4.2. *Weighting the actors and attributes*

One of the crucial tasks of this component is to determine the *initial* weights of the actors and the attributes that allow for a better classification of different abnormal behaviors. The *weights* refer to both the significance degree of actors (w_l^b), which will be used as part of the *value function* that will produce the risk measure of the *transaction*, and the significance of the attributes (s_l^m), which is used to assess the risk measure of the *actors* for a particular abnormal behavior category *l*. Note that the set of the actors and associated attributes are considered to be relevant indicators of abnormal behaviors by the experts and are determined during the storyboard process.

The weights are defined by experts subjectively, but there should be some objective methods helping the experts to make their decisions more consistent and reliable. There are several well-known methods for weighting such as Delphi method [35], Rank Order Centroid (ROC) [36] Ratio Method [37] and pairwise comparison method [38]. Delphi method is conducted by a director who is responsible from preparing a questioner, directing it to the members and has independent communication with each panel members. Panel members are anonymous to each other and they iteratively response the questioner with their own idea with their arguments. The director should process the gathered information at the end of the each iteration and make some filtering and consolidations. After the first iteration each member can see others' opinion and may write an opposite comment for it or change his own idea regarding to coming new information. The iterations continue until a consensus is reached. Delphi method can be used to reach a consensus for any decision but it can also be used for weighting the attributes efficiently. But there are some difficulties with Delphi method such as consuming relevantly more resources and time, requiring a dedicated and willing director and panel members to the usefulness of the method, etc. [35].

The Rank Order Centroid (ROC) Method is a simple way for calculating weights from the ranks of a number of items and based on the idea that the decision makers usually can rank the items with respect to their importance much more easily than giving weights to them. The calculation of the weight from the ranks is as the following formula: $W_i = (1/M)\sum_{n=1}^{M} 1/n$ where M is the number of items and W_i is the weight for ith item. For example for $M = 5$, the distribution of the weights would be 0.46, 0.26, 0.16, 0.09, 0.03. Although the method is very simple, the weights are highly dispersed and none of attributes may have the same weight at the end.

The Ratio Method is another method for calculating the weights from the ranks assigned by a decision maker. Firstly the decision maker determines the ranks of the factors. Then assign the weights between 10 and 100 with respect to the rank orders as multiples of 10. Finally the weights are normalized. This method is also very simple, but the weights any attributes cannot be the same like ROC method.

In our model, the initial weights are determined by using the binary pairwise comparison method of the analytical hierarchy process (AHP) [38]. Briefly, in this methodology, the experts are requested to complete a comparison matrix and specify which actors (or similarly attributes) are most significant for a particular abnormal behavior category. Later, the rank of each actor (or attribute) is determined, and the normalized rank values are assigned as the weights associated with the actors, i.e., w_l^b (or attributes, i.e., s_l^b). The experts (users) begin their analysis with the *initial* weights and fine tune these parameters using the developed framework, as depicted in Fig. 6. Note that, unlike the classical machine learning process where the experts provide a set of possible features and leave the rest (e.g., feature selection, feature weighting, parameter tuning, etc.) to the training process and interfere later once more for choosing the best classifier (note that usually this is also automatic in many applications and classification performance is utilized at this stage), in the case of the IML method, the experts interact in the training process with the machine. That is to say, the experts take the results as a feedback and revise their decisions including the reweighting of the features, feature combinations, relational features, etc. until they are satisfied with the results [2,5]. Here we have conducted two problems to solve:

(1) Reaching the consensus of the group for each iteration.
(2) As considering the accuracy of the system, deciding whether the series is well enough for stopping the iterations.

Reaching the consensus of a group of decision makers is the well-defined problem in the literature. The consensus is defined as a cooperative process in which a group of decision makers develops and agrees to support a decision in the best interest of the whole [39]. The first issue in this problem is to represent the expert's opinion as numbers and/or linguistic preferences. According to the domain and expert's choice, the linguistic, numeric or linguistic–numeric preferences can be used in the methods [40]. The second issue is to determine the consensus degree of the group and measure the linguistic distance of each individual from current consensus labels over the preferences [40]. A more recent study propose a new model in order to increase the level of agreement within the group by representing more granular format of the fuzzy preference relations [39]. Since the experts usually a group of experts initially presents inconsistencies in their opinion a consistency measure can be incorporated [41], the inconsistency becomes another issue that has been incorporated by some studies [42]. On the other hand, regarding measuring the consensus there are two common approaches in the literature. While the first one is focusing on the expert set; the second one considers changing the initial expert set and focuses on the alternative set [43]. All those studies, as well as Delphi [35], ROC [36], and Ratio Method [37] methods are proposing an approach to reach the consensus by a single or a group of experts so that the final decision would be used in a system or business, etc. In our case, actor, commodity and weight set, i.e., the *series* is produced from the storyboard with by a group of experts using focus group [44] discussions. And the actors and attributes are weighted by the experts by using binary pairwise comparison [38] method. Since we incorporated with IML approach, and the result of the series can be evaluated by only after analyzing the claims labeled as fraudulent by eFAD, it is highly possible to repeat the experiments of a series for many times. Thus, because of the time limitations we take the normalized average of the weights of the attributes and actors determined by the experts as a consensus of them for the iteration. Then the experts can have a chance to see the consequences of the series while analyzing the claims labeled as fraudulent by eFAD and make evaluation about how to change the series not only by means of weights, but also the expert set, i.e., actors and attributes, as well. Since there is a possibility to repeat decision process for several times, the time and the efficiency of the decision and consensus method are becoming more important. The main factors helping us to select binary pairwise comparison method against others, i.e., Delphi method, were the simplicity, consistency and quickness.

The second problem is when to decide stopping the iterations. According to our study, deciding to stop iterations is a cost saving issue as stated in Section 5.2, rather than a heuristic issue. So if the result is well-enough (the cost for detecting is considerably less than the amount of detected cases), then the iterations can be stopped for the relevant series.

4.3. Two-stage data warehouse

One of the reasons for the lack of proactive solutions in the literature is the high computational performance requirement. To overcome this problem, a two-stage data warehouse is constructed. The first stage is a traditional star-schema data warehouse, which is loaded with the data from the operational database of the insurance company via certain data cleansing algorithms in an extract, transform, and load (ETL) process. The second stage is extracted from the first stage and includes all potentially desirable indicator attributes (in terms of the three modes, *periodical*, *cumulative*, and *differential*) to ready them for use in risk calculations, i.e., $T_{i,j}^{m,b}$. Therefore, the second stage of the warehouse ensures a high computational performance and enables proactive calculation of the claim risks. The second component of the eFAD Suite depicted in Fig. 6 represents the two-stage data warehouse of the eFAD framework.

Because the attributes scores (i.e., $T_{i,j}^{m,b}$) actually represent the degree of anomaly of the relationship between actors and commodities, rather than using the nominal values of the attributes, the *z-scores* (standardized scores) are used in risk calculations while determining them. The *z*-scores are commonly used measures of the distance of the attribute values from the means, which are independent of the scales of the corresponding samples. Because the *z*-scores are relevant to the standard deviation and mean of the sample set, before calculating the *z*-scores, the actors are clustered to determine the most similar actors for comparison. The clustering process is performed using the expectation maximization (EM) method [45] with the attributes of the actors that represent their working place and capacity. Lastly, the resulting *z*-scores are standardized into a 1–100 scale. Table 2 tabulates the distribution of 138 attributes for different actor–commodity relations included in the second stage of the data warehouse (i.e., $m = 1\ldots M$, where $M = 138$). The type and number of attributes assessed with respect to the storyboards of each fraud type are described in Section 4.1.

4.4. Risk assessment engine

The risk assignment engine is a component of the eFAD framework that evaluates first the risk levels of the actors and then the medical claims based on the actors' risk levels. The types of abnormal behaviors (F_l), actors (A_i^b), commodities (C_j), attributes ($T_{i,j}^{m,b}$), weights of the actors (w_l^b), and weights of the attributes (s_l^m) are used as the inputs of this component. All of these parameters are determined as the result of the storyboard process as described earlier. Note that the actors, commodities, and attributes together with the weights of the actors and attributes form a vector that is referred to as a *series* for each abnormal behavior. The eFAD Suite Management Console enables domain experts to define new abnormal behaviors whenever needed and then set a combination of weighted actors and their weighted attributes as a series. Recall that the series that are successful based on experiments that measure their accuracy are set as active series to be used in the risk assignment engine.

The risk assignment engine utilizes these series first in order to calculate the *risks of actors* with respect to attributes representing their relationships with commodities that are involved in a medical claim. Next the *risk of the transaction* is determined by utilizing the risk of the actors that are calculated in the previous step. The risks of individual actors and of the transactions are assumed to be represented by additive value functions. As presented in Eq. (1), the risk of a particular actor ($G^l(A)_i^b$) for a particular abnormal behavior l is the weighted sum of the relevant attribute *z*-scores of the actors (i.e., $T_{i,j}^{m,b}$), where the weights are s_l^m. Note that because there might be multiple commodities involved in the medical claim, each attribute representing the actor–commodity relationship is calculated separately, and the maximum value over the number of commodities is utilized in the weighted sum calculations. In order to ease the understanding of the following equations for the readers, we will first revisit the notation that is developed earlier.

Notation:

MC_k refers to the kth transaction.
A_i^b refers to the ith entity of actor type b. Note that the index i, which is actually $i(b, k)$, refers to a *particular actor* of type b in transaction k.
C_j refers to the jth commodity.
F_l refers to the lth fraud type.
$T_{i,j}^{m,b}$ refers to the *z-score* of the mth attribute that measures the relationship of particular actor i (of type b) and commodity j. Note that these *z-scores* are calculated based on the historical data for each *metric* and *actor*.
s_l^m refers to the weight of the mth attribute in terms of the lth fraud type. Note that s_l^m's are determined from the story board process discussed above.
w_l^b refers to the weight of the bth actor type in terms of the lth fraud type. Note that w_l^b's are determined from the story board process discussed above.
$G^l(A)_i^b$ is the risk of the ith actor of type b (where i is actually $i(b,k)$ as discussed above) in terms of the lth fraud type.
R_k^l is the overall risk of the kth transaction, (i.e., MC_k) in terms of the lth fraud type.

Eq. (1) demonstrates how the actor risks, i.e., $G^l(A)_i^b$ is determined.

$$G^l(A)_i^b = \sum_{m=1}^{m} \max_j(s_l^m * T_{i,j}^{m,b}) \tag{1}$$

After the *actor risks* are determined, the weighted sum of the risks of individual actors yields the overall risk of the medical claim (i.e., MC_k) for the abnormal behavior l, i.e., R_k^l. Eq. (2) presents how R_k^l is calculated.

$$R_k^l = \sum_{b=1}^{\#actors} w_l^b * G^l(A_i^b) \tag{2}$$

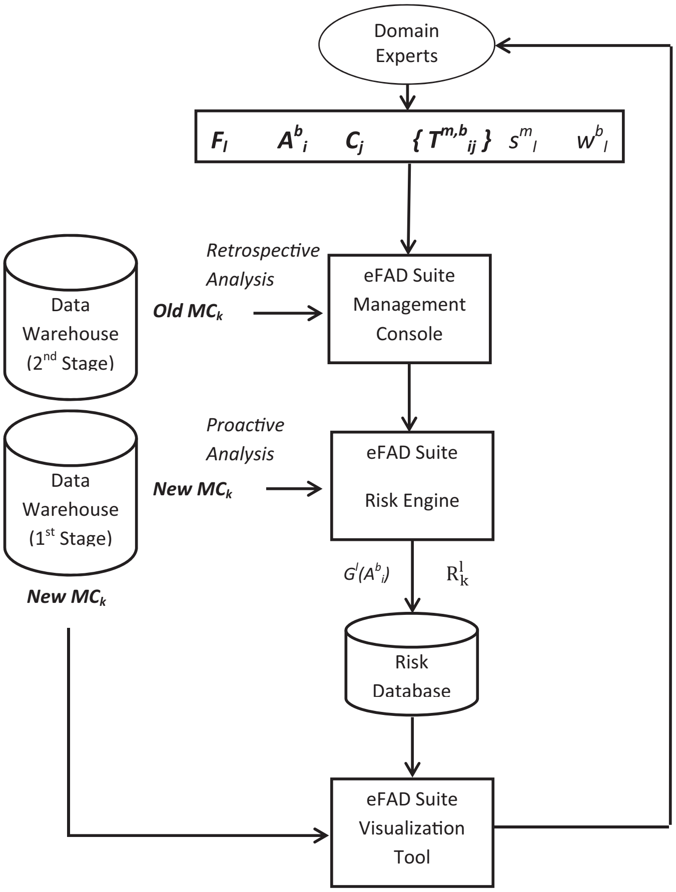

Domain Experts
F_l $A^b{}_i$ C_j $\{T^{m,b}{}_{ij}\}$ $s^m{}_l$ $w^b{}_l$
Retrospective Analysis
Old MC$_k$
Data Warehouse (2nd Stage)
eFAD Suite Management Console
Proactive Analysis
New MC$_k$
Data Warehouse (1st Stage)
eFAD Suite Risk Engine
New MC$_k$
$G^l(A^b{}_i)$
R^l_k
Risk Database
eFAD Suite Visualization Tool

Fig. 7. The incorporation of domain experts and basic inputs and outputs of the framework.

Please cite this article in press as: I. Kose, et al., An interactive machine-learning-based electronic fraud and abuse detection system in healthcare insurance, Appl. Soft Comput. J. (2015), http://dx.doi.org/10.1016/j.asoc.2015.07.018

Table 2
The number of attributes and types.

Actors	Commodities			
	Common	ATC code	Drug company	Guarantee type
Pharmacist	7	7	7	
Doctor	17	7	7	
Insured	29	7	7	5
Hospital	33			5

To determine the risk of all of the claims in terms of all possible fraud types, Eqs. (1) and (2) should be repeated accordingly.

4.5. Visualization tool

The first objective of the visualization component of the eFAD framework is fact-finding, i.e., the determination of the evidence (i.e., the rationale) for the decisions of the framework. Furthermore, it enables the experts (users) to learn more about the relationships among the actors, commodities, and abnormal behaviors so that they can revise their judgments regarding the parameters utilized during the training stage of the risk assessment engine, i.e., support the interactive machine-learning framework. The visualization tool is developed on the QlikView Business Intelligence Platform, which can be used for both reactive and proactive analysis. The proactive analysis focuses on the risk levels of the medical claims processed by the ECP system in real time with respect to various abnormal behavior types. Alternatively, the reactive analysis allows for users to conduct historical analysis to determine the risky actors and determine the reason for their high risk levels with respect to each abnormal behavior type. Fig. 7 represents the incorporation of the domain experts, the inputs and the outputs of the framework basically.

The proactive analysis provides a three-level visualization platform, as depicted in Figs. 8–11. The analysis begins with a radar chart (Fig. 8) of the most recent claims processed by the ECP, in which the risk levels of each claim are calculated using the eFAD framework, as described earlier. In the radar chart, the amplitude of the lines represents the risk levels of the individual claims. Once the user selects one of the most risky claims, the share of each actor in the composition of the corresponding risk level for the particular abnormal behavior is presented at the left-bottom of the screen. Next, the user chooses the actor he/she prefers to focus on, providing control charts for the relevant attributes of the chosen actor (Fig. 9). Note that the control chart is essentially a line chart that shows the *z-scores* of the particular attribute as a function of time together with the sample average and ± one standard deviation. This chart is interactive, and the user is able to focus on a particular period by selecting, as the third level of the visualization provides the list of the realized medical claims during that particular period (Fig. 10). The expert user can now investigate all transactions of the relevant actors in detail and determine whether the behavior is normal or abnormal with respect to medical and insurance point of view.

The reactive analysis also has three levels. The visualization tool starts with a 3-D bar chart that depicts the number of claims with abnormal behaviors for each actor with respect to each abnormal behavior type (Fig. 11). Note that a claim is classified as an abnormal behavior if it has a risk score over a predefined threshold value, and the 3-D bar chart depicts the number of such claims for each actor for each abnormal behavior type. This allows the user to easily detect the actors that are potentially the most risky. After the user chooses a particular actor and abnormal behavior type combination, the visualization tool directs the user to the second level, which presents a control chart illustrating the $T_{i,j}^{m,b}$ of the relevant attributes. Similarly, the third level shows the realized medical claims for a period of time that is specified by the user.

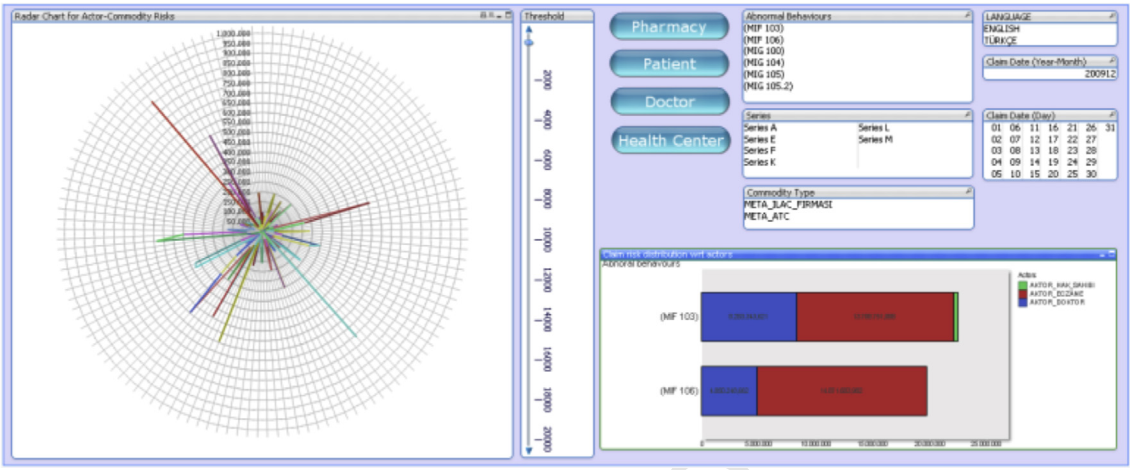
5. Results

An experimental analysis is conducted to measure the accuracy of eFAD. The experiments use real-life data from one of the leading insurance companies in Turkey (Company ABC), providing health insurance to ≈100,000 individuals. The data set consists of 845,247 prescription claims collected over four years, namely, 2008–2011.

Recall that the actors, attributes, and associated weights were determined earlier via the storyboards by the experts and deposited in the eFAD Suite Management Console as *series*. First, the risks of all 845,247 prescriptions were calculated using eFAD with respect to the series of each abnormal behavior type. Those risk values were taken as *predictions* according to whether they were higher than the preset threshold (here, average + 1 standard deviation). Second, the subsets, including 48 sample prescriptions, were selected randomly for each six abnormal behavior types (i.e., the total number of claims randomly selected is equal to 288). Recall that we used three types of attributes such as periodical, differential and cumulative as explained in Section 4.1. Since all attributes represent the *z*-score of the relation of the actors and commodities, we can differentiate the long term and short term fraudulent behaviors from each other by selecting the appropriate attribute types. For example, while periodical or differential attribute types are used in Series A, we prefer to use cumulative attribute types in Series B with different weights. Finally, the names of the series listed in Tables 3 and 4 and Fig. 12 are as follows:

Series A: Prescribing abnormally of a specific drug company.
Series B: Prescribing abnormally of a specific drug company during the last year.
Series C: Prescribing abnormally the costly drugs of a specific ATC code.
Series D: Prescribing abnormally the costly drugs of a specific ATC code during the last year.
Series E: Prescribing abnormally the drugs of a specific ATC code.
Series F: Prescribing abnormally the drugs of a specific ATC code during the last year.

Although the risk calculation was performed on whole dataset, because of the cost of expert evaluation, we aimed to select a representative subset for each abnormal behavior types. Third, the experts were asked to evaluate the selected claims (prescriptions) regardless of whether these claims were definitely the *part* of the predicted abnormal behavior type by eFAD. The experts' decisions were taken as the *actual* status of the claims. It is important to note that the experts were only informed about the abnormal behavior type for each subset to evaluate, not the calculated risks of the claims. Recall that most frauds cannot be determined individually due to the fragmented nature of such behaviors. Therefore, in this process, the experts utilized the *visualization tool* to make the assessment. However, they were not exposed to the radar chart of the visualization tool in Fig. 8 to prevent them from seeing the risks of the claims and were directed to the control charts as depicted in



Abnormal Behaviours
(MIF 103)
(MIF 106)
(MIG 100)
(MIG 104)
(MIG 105)
(MIG 105.2)

Series
Series A
Series E
Series F
Series K
Series L
Series M

Commodity Type
META_ILAC_FIRMASI
META_ATC

LANGUAGE
ENGLISH
TÜRKÇE

Claim Date (Year-Month)
200912

Claim Date (Day)
01 06 11 16 21 26 31
02 07 12 17 22 27
03 08 13 18 23 28
04 09 14 19 24 29
05 10 15 20 25 30

Fig. 8. The radar chart of proactive analysis modules of eFAD suite.

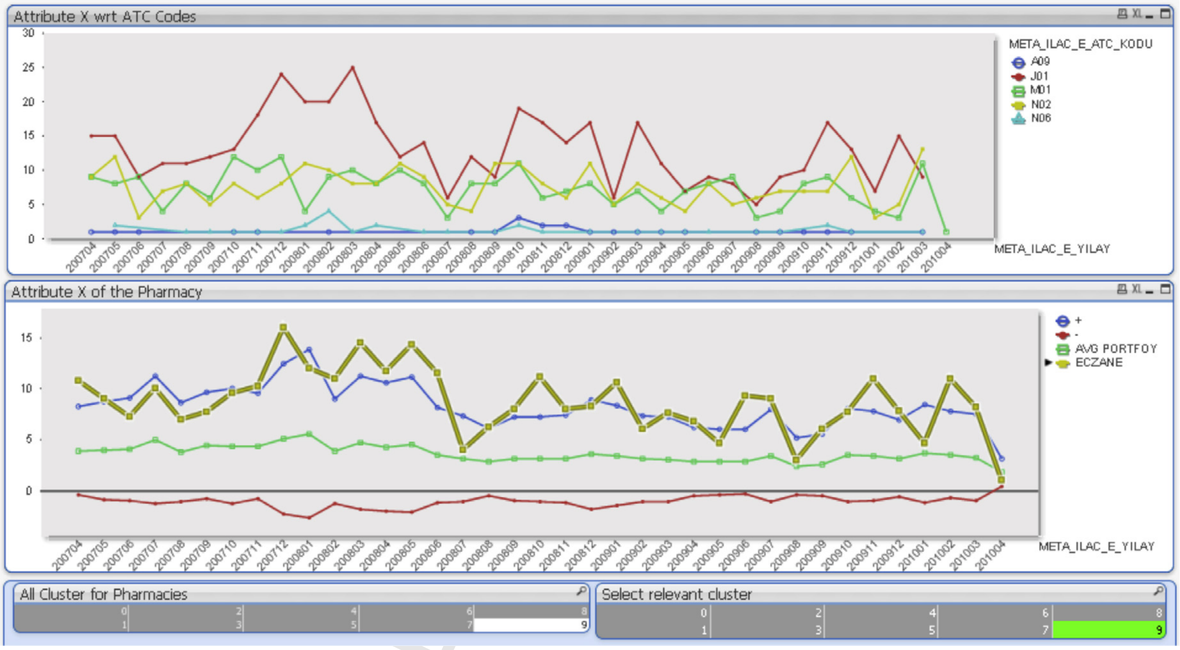
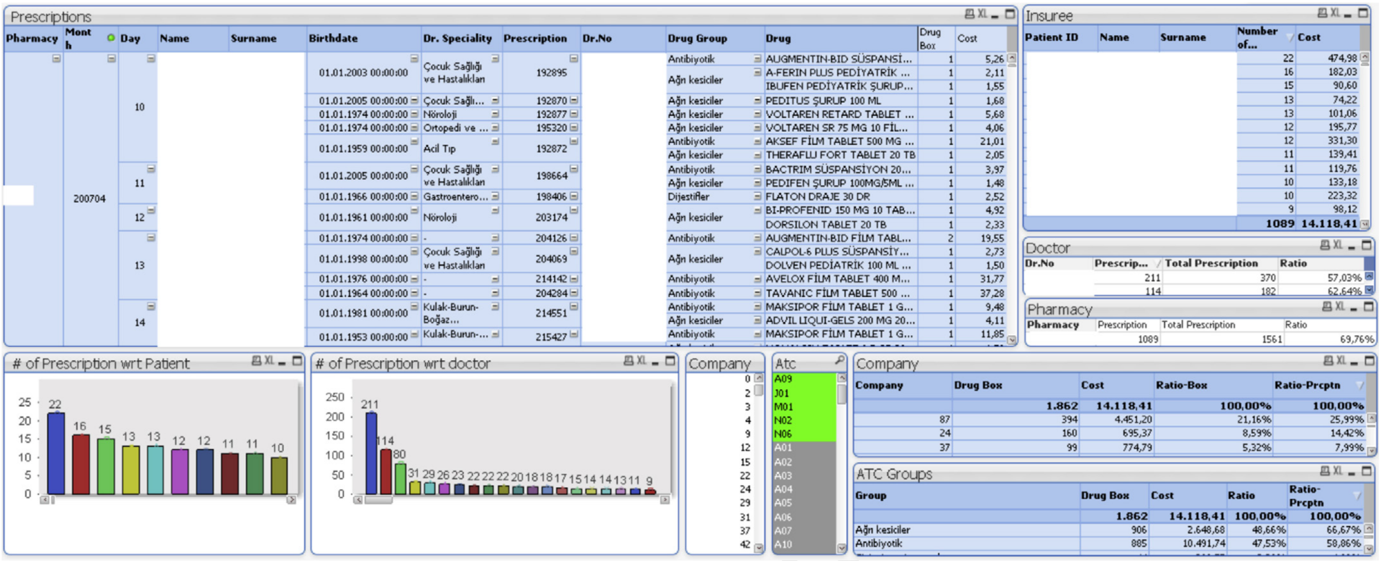


Fig. 9. The control charts depicting the changing of attributes of actor among the relevant actor cluster.

Table 3
The confusion matrix of ZeroR classifier for experimented series.

Abnormal behavior	Actual	Prediction		Accuracy
		Positive	Negative	
A	Positive	28	20	58.3%
	Negative	0	0	
B	Positive	31	17	64.5%
	Negative	0	0	
C	Positive	24	24	50%
	Negative	0	0	
D	Positive	22	26	45.8%
	Negative	0	0	
E	Positive	28	20	58.3%
	Negative	0	0	
F	Positive	26	22	54.1%
	Negative	28	20	

Please cite this article in press as: I. Kose, et al., An interactive machine-learning-based electronic fraud and abuse detection system in healthcare insurance, Appl. Soft Comput. J. (2015), http://dx.doi.org/10.1016/j.asoc.2015.07.018



Prescriptions

Pharmacy	Month	Day	Name	Surname	Birthdate	Dr. Speciality	Prescription	Dr.No	Drug Group	Drug	Drug Box	Cost
	200704	10			01.01.2003 00:00:00	Çocuk Sağlığı ve Hastalıklan	192895		Antibiyotik	AUGMENTIN-BID SÜSPANSİ...	1	5,26
									Ağn kesiciler	A-FERIN PLUS PEDİYATRİK ...	1	2,11
										IBUFEN PEDİYATRİK ŞURUP...	1	1,55
					01.01.2005 00:00:00	Çocuk Sağlı...	192870		Ağn kesiciler	PEDITUS ŞURUP 100 ML	1	1,68
					01.01.1974 00:00:00	Nöroloji	192877		Ağn kesiciler	VOLTAREN RETARD TABLET ...	1	5,68
					01.01.1974 00:00:00	Ortopedi ve ...	195320		Ağn kesiciler	VOLTAREN SR 75 MG 10 FİL...	1	4,06
					01.01.1959 00:00:00	Acil Tıp	192872		Antibiyotik	AKSEF FİLM TABLET 500 MG ...	1	21,01
									Ağn kesiciler	THERAFLU FORT TABLET 20 TB	1	2,05
		11			01.01.2005 00:00:00	Çocuk Sağlığı ve Hastalıklan	198664		Antibiyotik	BACTRIM SÜSPANSİYON 20...	1	3,97
									Ağn kesiciler	PEDIFEN ŞURUP 100MG/5ML ...	1	1,48
					01.01.1966 00:00:00	Gastroentero...	198406		Dijestifler	FLATON DRAJE 30 DR	1	2,52
		12			01.01.1961 00:00:00	Nöroloji	203174		Ağn kesiciler	BI-PROFENID 150 MG 10 TAB...	1	4,92
										DORSILON TABLET 20 TB	1	2,33
		13			01.01.1974 00:00:00	-	204126		Antibiyotik	AUGMENTIN-BID FİLM TABL...	2	19,55
					01.01.1998 00:00:00	Çocuk Sağlığı ve Hastalıklan	204069		Ağn kesiciler	CALPOL-6 PLUS SÜSPANSİY...	1	2,73
										DOLVEN PEDİATRİK 100 ML ...	1	1,50
					01.01.1976 00:00:00	-	214142		Antibiyotik	AVELOX FİLM TABLET 400 M...	1	31,77
					01.01.1964 00:00:00	-	204284		Antibiyotik	TAVANIC FİLM TABLET 500 ...	1	37,28
		14			01.01.1981 00:00:00	Kulak-Burun-Boğaz...	214551		Antibiyotik	MAKSIPOR FİLM TABLET 1 G...	1	9,48
									Ağn kesiciler	ADVIL LIQUI-GELS 200 MG 20...	1	4,11
					01.01.1953 00:00:00	Kulak-Burun-...	215427		Antibiyotik	MAKSIPOR FİLM TABLET 1 G...	1	11,85

Insuree

Patient ID	Name	Surname	Number of...	Cost
			22	474,98
			16	182,03
			15	90,60
			13	74,22
			13	101,06
			12	195,77
			12	331,30
			11	139,41
			11	119,76
			10	133,18
			10	223,32
			9	98,12
			1089	14.118,41

Doctor

Dr.No	Prescrip...	Total Prescription	Ratio
	211	370	57,03%
	114	182	62,64%

Pharmacy

Pharmacy	Prescription	Total Prescription	Ratio
	1089	1561	69,76%

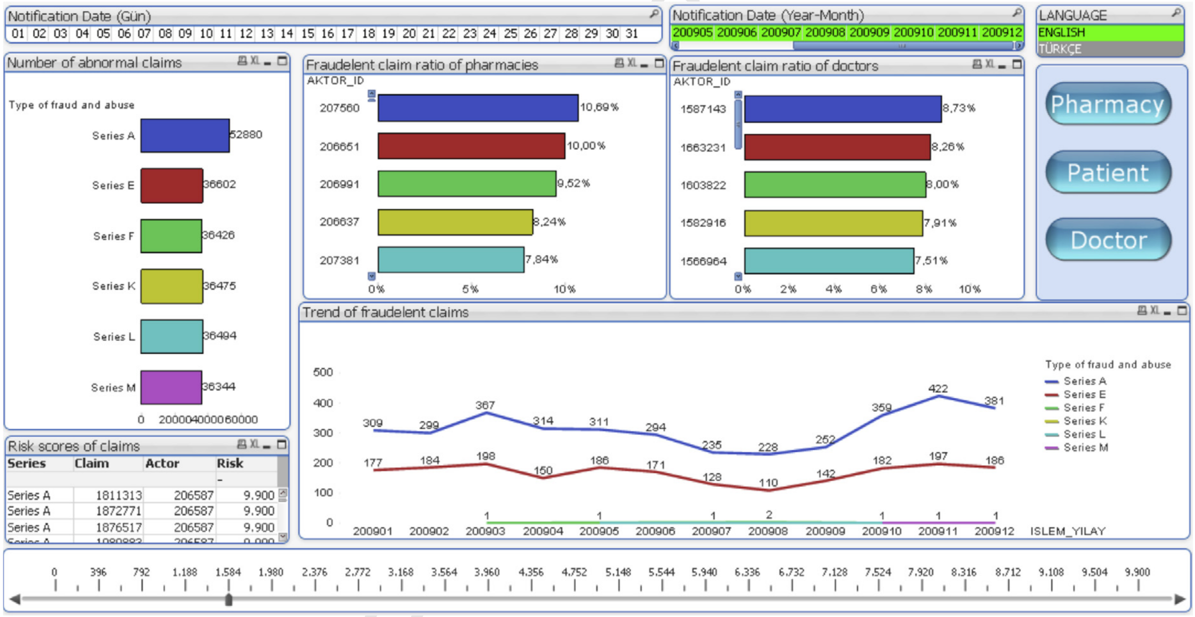
Company

Company	Drug Box	Cost	Ratio-Box	Ratio-Prcptn
	1.862	14.118,41	100,00%	100,00%
87	394	4.451,20	21,16%	25,99%
24	160	695,37	8,59%	14,42%
37	99	774,79	5,32%	7,99%

ATC Groups

Group	Drug Box	Cost	Ratio	Ratio-Prcptn
	1.862	14.118,41	100,00%	100,00%
Ağn kesiciler	906	2.648,68	48,66%	66,67%
Antibiyotik	885	10.491,74	47,53%	58,86%

Fig. 10. Screenshots of proactive and reactive modules of eFAD suite.



Risk scores of claims

Series	Claim	Actor	Risk
			-
Series A	1811313	206587	9.900
Series A	1872771	206587	9.900
Series A	1876517	206587	9.900

Fig. 11. The screenshot of a form of reactive analysis module depicting the risky claims distribution according to actors and fraud types.

Table 4
The accuracy and AUC of eFAD and accuracy of ZeroR classifier of several experimented abnormal behaviors (for threshold = mean + standard deviation).

Abnormal Behavior	Actual	Prediction		Accuracy (eFAD)	Accuracy (ZeroR)	AUC (eFAD)
		Positive	Negative			
A	Positive	89.3%	10.7%	79.2%	58.3%	82.1%
	Negative	35.0%	65.0%			
B	Positive	93.5%	6.5%	89.6%	64.5%	92.2%
	Negative	17.6%	82.4%			
C	Positive	91.7%	8.3%	75.00%	50%	83.3%
	Negative	41.7%	58.3%			
D	Positive	90.9%	9.1%	70.80%	45.8%	75.2%
	Negative	46.2%	53.8%			
E	Positive	89.3%	10.7%	79.20%	58.3%	87.1%
	Negative	35.0%	65.0%			
F	Positive	88.5%	11.5%	75.00%	54.1%	86.4%
	Negative	40.9%	59.1%			

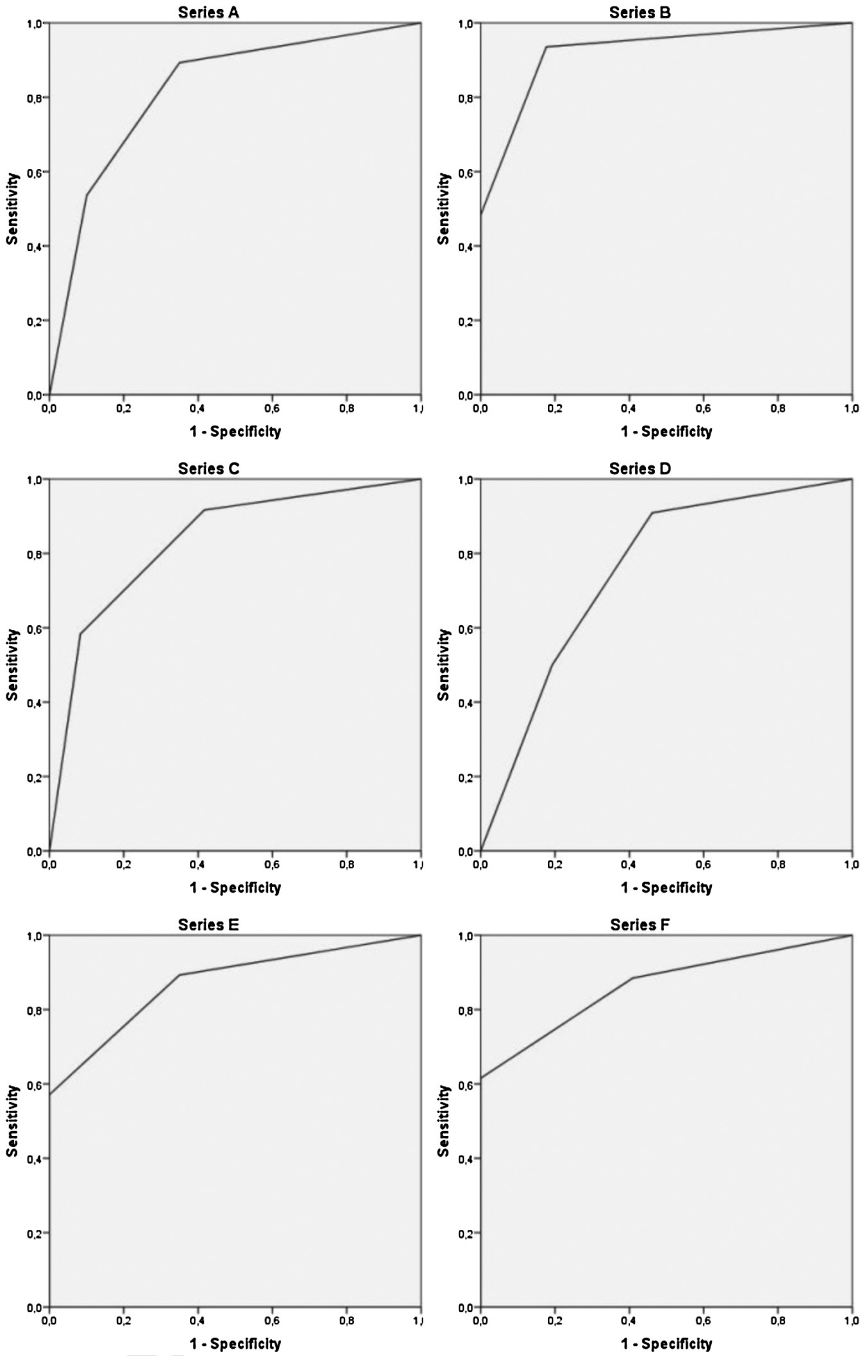


Fig. 12. ROCs of experimental analysis.

Figs. 9–11 in order to analyze the relationships between the actors and commodities in terms of the relevant attributes defined in the series.

Because the IML enables experts to choose certain parameters, evaluate and compare models, and choose those that are most appropriate for their goals, the experts used the results of the experiments to revise the contents of the series. Further iterations for the same series were then executed until the accuracy no longer changed meaningfully.

5.1. The accuracy model

The *predictions* in the accuracy model are defined by the risk values of the claims relative to a threshold. The *actuals* are the final decisions of the experts based on their investigation of all of the relevant information using the visualization tool. We should note that, since the attributes, data sets and claim types differ in all studies; there is no any benchmark for this problem to compare with our study. Thus the ZeroR classifier is incorporated in the model as a baseline as tabulated in Table 3.

The accuracy of the eFAD framework is also tabulated in Table 4 in terms of the true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) for a threshold level of the mean + one standard deviation for the six abnormal behavior types. The final column in the table presents the AUC of the receiver operating characteristic curves (ROC) obtained by changing the threshold levels. The ROC curves are also presented in Fig. 12.

Table 5
The cost of the system outcomes.

Outcome	Cost
Cost of true positives	# of TPs × average cost per investigation
Cost of true negatives	# of TNs × average cost per claim
Cost of false positives	# of FPs × (average cost per investigation + average cost per claim)
Cost of false negatives	# of FNs × average cost per claim

Table 4 demonstrates that the true positive rates (i.e., hits) vary between 88.5% and 93.5%. These numbers reveal that the *sensitivity* of the proposed framework is very high and that the model can detect most of the abnormal behaviors in the data set. Alternatively, the *specificity*, i.e., the true negative rates, varies between 53.8% and 82.4%. The high variability in terms of the specificity of the results for different abnormal behaviors might hint that the threshold level of *plus one standard deviation* might not be a good choice for certain fraud types. However, as the overall accuracy levels indicate (which vary between 70.8% and 89.6% for various abnormal behaviors), the performance of the proposed framework is encouraging. On the other hand the AUC values of the ROC curves vary between 75.2% and 92.2%.

5.2. *The cost-savings model*

Cost savings is another indicator to measure the benefit of a system. The introduced cost-savings model [46] assumes that all alerts must be investigated. Additionally, the average cost per claim must be higher than the average cost per investigation. Table 5 illustrates the costs that would be incurred for different system outputs.

Note that there are two extreme situations. First, one can execute all of the claims as normal (that is, take *no action*). The second extreme is the case in which the model accuracy is perfect, i.e., TP and TN are 100% (the best case). However, in reality, the system will perform somewhere between these two extremes in terms of cost savings. Therefore, the following evaluation metrics are utilized in the cost-savings model:

$$\text{Model cost savings} = \text{No action} - [\text{costs of } (TP + TN + FP + FN)] \tag{3}$$

$$\text{Percentage saved} = \left(\frac{\text{Model cost savings}}{\text{Best case scenario cost savings}} \times 100\right) \tag{4}$$

We conducted a cost-savings analysis for a more general context, e.g., a province or country, as opposed to a limited version of such an analysis for a particular private insurance company. Hence, to calculate the *model cost savings*, the typical average claim cost in Turkey (or a relevant approximation) is required for the analysis. Mollahaliloglu et al. [47] reported that the *average cost per prescription* in Denizli (a province in Turkey with a population of nearly 1 million) as 133 TL (as of April, 2012; ≈73 USD). Alternatively, to determine the *investigation cost*, a time study lasting for three days is conducted, revealing that this value is approximately 9.5 USD per claim. Based on these cost figures, with an assumption of 10% abnormal behavior in all claims, as suggested in the National Health Care Anti-Fraud Association (1991) report and a typical accuracy of 80% for the proposed framework, a projected cost-savings analysis is conducted for Denizli province.

The resulting percentage saved is 4.71%. This figure implies that the cost savings for Denizli province alone would be approximately 11.7 million USD if the officials utilized the proposed framework. Note that this number is obtained from a projection of 273 million USD for the total prescription damage, which assumes that the total prescription damage for the province is proportional to its share of the country's population. Alternatively, the best-case scenario would provide a 9.53% savings with 23.7 million USD in the case of 100% TP or TN (i.e., 100% accuracy).

6. Conclusions

In this research, we developed a novel interactive machine-learning-based framework for detecting fraud and abusive behaviors in the healthcare insurance industry. The developed framework is actor- and commodity-independent, configurable (i.e., easily adaptable in the dynamic environment of fraud and abusive behaviors), and effectively handles the *fragmented* nature of the abnormal behaviors. It can be used both for proactive and reactive analyses. The framework also includes a visualization tool that significantly reduces the time requirements for the users during the fact-finding process after the eFAD Suite alerted the user of risky claims.

The developed framework is among the few successful applications of fraud and abusive behavior detection in healthcare insurance. Recall that the literature consists of applications that limit their focus to specific actors, such as physicians, patients, or healthcare providers, or distantly related applications, such as health report fraud or clinical pathways. Therefore, the developed framework is the only application that includes nearly all of the actors and commodities in the healthcare insurance domain and provides a solution that can satisfy the demand for a decision support tool that can assign risk measures to claim transactions.

The performance of the proposed framework is assessed using experimental analysis. By incorporating with the experts' experience, the storyboards are prepared and relevant actors and attributes are selected regarding the storyboards of the fraud types by a focus groups study. Then experts are requested to express their opinion about weighting by using pairwise comparison method. The normalized average weights of the experts are taken as the consensus of the experts for the experimental iterations. Until the results of the experiments are satisfactory in terms of both accuracy (i.e., the AUC of ROC curves) and cost savings model, the experimental iterations continued. As a result, the proposed framework is developed and marketed as eFAD Suite as a part of a dynamic damage management service provided by CGM Turkiye for private insurance companies.

Meanwhile, this system is used by CGM on behalf of its customers to produce monthly reports, including abnormal behavior cases to be evaluated and acted upon by the insurance company.

Although it is specifically developed for healthcare insurance systems, this framework can also be adapted to other insurance contexts due to the flexibility of its actor–commodity-based approach. This task is left as a further research topic. On the other hand, a thorough analysis of the interactive machine learning process, particularly the details of the experts' opinion revision process (e.g., what are the factors and how they influence the decision making process, is it purely a trial-error or a structured learning process or a hybrid version adopted, etc.) will be valuable for the future IML applications. Additionally, we should note that most of the attributes may be used within some other series to detect different fraud types. Thus, when an expert changes the weight of an attribute, then the series may become similar to another series of another fraud type and start to catch some claims which may belong to this relevant fraud type. Besides, the set of frauds is dynamic, and the predefined fraud types and their series including actors, commodities, attributes and weights should change in time or new fraud types should be defined.

Please cite this article in press as: I. Kose, et al., An interactive machine-learning-based electronic fraud and abuse detection system in healthcare insurance, Appl. Soft Comput. J. (2015), http://dx.doi.org/10.1016/j.asoc.2015.07.018

Lastly, incorporating an engine that also considers the relationships between actors, i.e., the network risk of the actors (i.e., actors that are related to more risky actors become more risky themselves and those related to less risky actors become less risky themselves) is also being considered as a further extension of the framework.

Acknowledgments

This work was supported in part by The Scientific and Technological Research Council of Turkey (TUBITAK) under project (grant) number 3090145 (TEYDEB). The authors thank Prof. Aytul Ercil and Associate Profs. Yucel Saygin and Muratcan Cobanoglu from Sabanci University as well as Fatma Zakaryan, M.D., General Manager of CGM Turkey, for all of their support and valuable collaborations. The authors are also grateful to the medical and insurance experts Omur Sezer, Dilek Pekiner, M.D., Ibrahim Aktas, M.D., Adil Karasoy, M.D., Ulas Ozcitak, M.D., and Emine Emre for their medical and insurance expertise and Mehmet Bilcin, Murat Metin, Emin Tuyluce, Nagehan Acer, and Pinar Horuz for their valuable technical expertise and efforts.

References

[1] R. Porter, J. Theiler, D. Hush, Interactive Machine Learning in Data Exploitation, No. LA-UR-13-20441, Los Alamos National Laboratory (LANL), 2013.

[2] J.A. Fails, D.R. Olsen Jr., Interactive machine learning, in: Proc. of the 8th International Conference on Intelligent User Interfaces, ACM, 2003, pp. 39–45.

[3] A. Holzinger, I. Jurisica, Knowledge discovery and data mining in biomedical informatics: the future is in integrative, interactive machine learning solutions, in: Interactive Knowledge Discovery and Data Mining in Biomedical Informatics, Springer Berlin Heidelberg, 2014, pp. 1–18.

[4] M. Ware, E. Frank, G. Holmes, M. Hall, I.H. Witten, Interactive machine learning: letting users build classifiers, Int. J. Hum.-Comput. Stud. 55 (2001) 281–292.

[5] S. Stumpf, V. Rajaram, L. Li, W.K. Wong, M. Burnett, T. Dietterich, E. Sullivan, J. Herlocker, Interacting meaningfully with machine learning systems: three experiments, Int. J. Hum.-Comput. Stud. 67 (2009) 639–662.

[6] C. Turkay, F. Jeanquartier, A. Holzinger, H. Hauser, On computationally-enhanced visual analysis of heterogeneous data and its application in biomedical informatics, in: Interactive Knowledge Discovery and Data Mining in Biomedical Informatics, Springer Berlin Heidelberg, 2014, pp. 117–140.

[7] A. Holzinger, Human–computer interaction and knowledge discovery (HCI-KDD): what is the benefit of bringing those two fields to work together? in: A. Cuzzocrea, C. Kittl, D.E. Simos, E. Weippl, L. Xu (Eds.), CD-ARES 2013, LNCS, Vol. 8127, Springer, Heidelberg, 2013, pp. 319–328.

[8] R. Fiebrink, P.R. Cook, D. Trueman, Human model evaluation in interactive supervised learning, in: CHI 2011, Session: Machine Learning, Vancouver, BC, Canada, May, 2011.

[9] V. Chandola, A. Banerjee, V. Kumar, Anomaly detection: a survey ACM Comput. Surv. 41 (July) (2009) 15.

[10] W.S. Yang, A Process Pattern Mining Framework for the Detection of Health Care Fraud and Abuse (Ph.D. thesis), National Sun Yat-Sen University, Taiwan, 2003.

[11] R.J. Bolton, D.J. Hand, Statistical fraud detection: a review, Stat. Sci. 17 (2002) 235–255.

[12] Centers for Medicare and Medicaid Services, National Health Expenditures Aggregate, Per Capita Amounts, Percent Distribution, and Average Annual Percent Change: Selected Calendar Years 1960–2010, Table 1, 2010, Retrieved from: https://www.cms.gov/NationalHealthExpendData/downloads/tables.pdf

[13] M. Kumar, R. Ghani, Z.S. Mei, Data mining to predict and prevent errors in health insurance claims processing, in: KDD'10, Washington, DC, USA, 2010.

[14] The Electronic Claim Processing for Drug (ECP Drug) product of CGM, Retrieved May 2014 from: http://www.cgm.com/tr/products___solutions_3/health_payers/cgm_ecp_drug/cgm_ecp_drug.en.jsp.

[15] W.S. Yang, S.Y. Hwang, A process mining framework for the detection of healthcare fraud and abuse, Expert Syst. Appl. 31 (2006) 56–68.

[16] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, From data mining to knowledge discovery in databases, Am. Assoc. Artif. Intell. Fall (1996) 37–54.

[17] National Health Care Anti-Fraud Association (NHCAA), Guidelines to Health Care Fraud: Factsheet, 1991, Retrieved May 2002 from: http://www.nhcaa.org

[18] U.S. Federal Bureau of Investigation, Financial Crimes Report to the Public, 2007, Retrieved May 2012 from: http://www.fbi.gov/stats-services/publications/fcs_report 2007

[19] L. Copeland, Applying business intelligence concepts to medicaid claim fraud detection, in: Conference for Information Systems Applied Research, CONISAR Proceedings, Wilmington North Carolina, USA, 2011.

[20] R.A. Derrig, Insurance fraud, J. Risk Insur. 69 (2002) 271–287.

[21] D. Castro, Improving Health Care Why a Dose of it May Be Just What the Doctor Ordered, The Information Technology & Innovation Foundation, 2007, pp. 1–23.

[22] C. Plaisant, B. Shneiderman, An information architecture to support the visualization of personal histories, Inf. Process. Manag. 34 (1998) 581–597.

[23] National Health Care Anti-Fraud Association (NHCAA), The Problem of Health Care Fraud, 2002, Retrieved May 2012 from: http://www.nhcaa.org/eweb/DynamicPage.aspx?webcode=anti_fraud_resource_centr&wpscode=TheProblemOfHCFraud

[24] C. Phua, V. Lee, K. Smith, R.A. Gayler, Comprehensive Survey of Data Mining-Based Fraud Detection Research in Technical Report, Monash University, 2005.

[25] E.W.T. Ngai, H. Yong, Y.H. Wong, Y. Chen, X. Sun, The application of data mining techniques in financial fraud detection: a classification framework and an academic review of literature, Decis. Support Syst. J. 50 (2011) 559–569.

[26] J. Li, K.Y. Huang, J. Jin, J. Shi, A survey on statistical methods for health care fraud detection, Health Care Manag. Sci. 11 (2008) 275–287.

[27] H. He, J. Wang, W. Graco, S. Hawkins, Application of neural networks to detection of medical fraud, Expert Syst. Appl. 13 (1997) 329–336.

[28] G. Williams, Evolutionary hot spots data mining: an architecture for exploring for interesting discoveries, in: Proc. 3rd Pacific-Asia Conference on Knowledge Discovery and Data Mining, Japan, 1999.

[29] K. Yamanishi, J.I. Takeuchi, On-line unsupervised outlier detection using finite mixture with discounting learning algorithms, Data Min. Knowl. Discov. 8 (2004) 275–300.

[30] J.A. Major, R. Dan, Riedinger, EFD: a hybrid knowledge/statistical-based system for the detection of fraud, J. Risk Insur. 69 (2002) 309–324.

[31] P.A. Ortega, C.J. Figueroa, G.A. Ruz, A medical claim fraud/abuse detection system based on data mining: a case study in Chile, in: Proc. The International Conference on Data Mining (DMIN), Las Vegas, Nevada, USA, June, 2006.

[32] L. Sokol, B. Garcia, J. Rodriguez, M. West, K. Johnson, Using data mining to find fraud in HCFA health care claims, Top. Health Inf. Manag. 22 (2001) 1–13.

[33] K.D. Aral, H.A. Guvenir, İ. Sabuncuoglu, A.R. Akar, A prescription fraud detection model, Comput. Methods Programs Biomed. 106 (2012) 37–46.

[34] M.E. Johnson, N. Nagarur, Multi-stage methodology to detect health insurance claim fraud, Health Care Manag. Sci. (2015), http://dx.doi.org/10.1007/s10729-015-9317-3

[35] A.P.C. Chan, E.H.K. Yung, P.T.I. Lam, C.M. Tam, S.O. Cheung, Application of Delphi method in selection of procurement systems for construction projects, Constr. Manag. Econ. 19 (7) (2001) 699–718.

[36] W. Edwards, F.H. Barron, SMARTS and SMARTER: improved simple methods for multiattribute utility measurement, Organ. Behav. Hum. Decis. Process. 60 (3) (1994) 306–325.

[37] M. Weber, K. Borcherding, Behavioral influences on weight judgments in multiattribute decision making, Eur. J. Oper. Res. 67 (1) (1993) 1–12.

[38] H. Taira, Y. Fan, K. Yoshiya, H. Miyagi, A method of constructing pairwise comparison matrix in decision making, in: Proc. Systems, Man, and Cybernetics, IEEE International Conference, vol. 4, 1996, pp. 2511–2516.

[39] F.J. Cabrerizo, R. Ureña, W. Pedrycz, E. Herrera-Viedmab, Building consensus in group decision making with an allocation of information granularity, Fuzzy Sets Syst. 255 (2014) 115–127.

[40] F. Herrera, E. Herrera-Viedma, J.L. Verdegay, A rational consensus model in group decision making using linguistic assessments, Fuzzy Sets Syst. 88 (1997) 31–49.

[41] F. Herrera, E. Herrera-Viedma, J.L. Verdegay, A model of consensus in group decision making under linguistic assessments, Fuzzy Sets Syst. 78 (1996) 73–87.

[42] Z. Wu, J. Xu, A consistency and consensus based decision support model for group decision making with multiplicative preference relations, Decis. Support Syst. 52 (2012) 757–767.

[43] E. Herrera-Viedma, F.J. Cabrerizo, J. Kacprzyk, W. Pedrycz, A review of soft consensus models in a fuzzy environment, Inf. Fusion 17 (2014) 4–13.

[44] R.A. Krueger, M.A. Casey, Focus Groups, third ed., Sage Publications, Thousand Oaks, CA, 2000.

[45] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, J. R. Stat. Soc. B 39 (1) (1977) 1–38.

[46] C. Phua, D. Alahakoon, V. Lee, Minority report in fraud detection: classification of skewed data, SIGKDD Explor. 6 (2004) 50–59.

[47] S. Mollahaliloglu, A. Alkan, B. Donertas, S. Ozgulcu, A. Akici, Assessment of the prescriptions written in different provinces of turkey in terms of drug utilization principles, Marmara Med. J. 24 (2011) 162–173.

Please cite this article in press as: I. Kose, et al., An interactive machine-learning-based electronic fraud and abuse detection system in healthcare insurance, Appl. Soft Comput. J. (2015), http://dx.doi.org/10.1016/j.asoc.2015.07.018