Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2012, Article ID 403987, 11 pages
doi:10.1155/2012/403987

*Research Article*

# Using Medical History Embedded in Biometrics Medical Card for User Identity Authentication: Privacy Preserving Authentication Model by Features Matching

**Simon Fong and Yan Zhuang**

*Department of Computer and Information Science, University of Macau, Taipa, Macau*

Correspondence should be addressed to Simon Fong, ccfong@umac.mo

Received 20 December 2011; Accepted 25 December 2011

Academic Editor: Sabah Mohammed



Many forms of biometrics have been proposed and studied for biometrics authentication. Recently researchers are looking into longitudinal pattern matching that based on more than just a singular biometrics; data from user's activities are used to characterise the identity of a user. In this paper we advocate a novel type of authentication by using a user's medical history which can be electronically stored in a biometric security card. This is a sequel paper from our previous work about defining abstract format of medical data to be queried and tested upon authentication. The challenge to overcome is preserving the user's privacy by choosing only the useful features from the medical data for use in authentication. The features should contain less sensitive elements and they are implicitly related to the target illness. Therefore exchanging questions and answers about a few carefully chosen features in an open channel would not easily or directly expose the illness, but yet it can verify by inference whether the user has a record of it stored in his smart card. The design of a privacy preserving model by backward inference is introduced in this paper. Some live medical data are used in experiments for validation and demonstration.

## 1. Introduction

The latest trend in biometrics authentication nowadays is to use multiple biometrics [1, 2] for extra security and users' longitudinal activity patters for identifying the users. The latter one appears to be appealing because it is generally more difficult to erase or forge a full history record about a person as history involves event records in multiple parties over a long period of time. Recently some advances in biometrics theories are based on one's email history patterns, online activity log patterns, and other personal history events [3, 4]. In this paper, we advocate the use of medical history data as biometrics as they may equally well in distinguishing a person and they are not easily counterfeited. Each medical record is handled supposedly by licensed medical professional (compare to a log on email file server or other public online platforms), hence medical records should be quite credible. Two adults are hardly having exactly the same medical history in terms of conditions, prognosis, treatment procedures, times, and places over a certain length of time.

One of the major challenges in using medical history for authentication, however, is privacy issue. Humans are naturally reluctant to reveal their private medical records and they may feel inferior if such data are openly communicated in the public for authentication. As shown in Figure 1, the medical history data that is stored in a smart card could be used for both medical consultations in different clinics where they may not be able to access a common patients' database and for authentication in addition to passwords or other forms of biometrics like fingerprints and iris scan. The authenticator in this case may be a machine device or a human officer that is able to generate some question-and-answer type of challenges to the testing user about his medical history. Only the authentic user is supposed to possess the knowledge of his own medical history, and he would be able to correctly answer the questions.

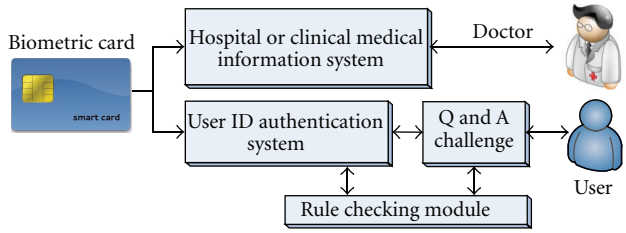



Figure 1: Workflow of the two uses of medical history from the biometric card.

Due to privacy reason, the questions to be asked should not come directly from the medical conditions. Embarrassments often occur especially when sensitive questions were raised in public about the users, for example, "Did you ever have a nose implant before? Did you start to suffer from erectile dysfunction last week?" A self-help authentication may avoid the embarrassment. However, it is impractical too to assume some costly machines with I/O devices are always available and proliferate everywhere, which can silently generate and display the questions on a small screen and receive input from the users. At times human officers are deployed and such questions may be asked in person in a public place. Given this privacy issue, a privacy-preserving mechanism is therefore much desired so that questions will not be directly asked from the medical illness but authentication by referring to the medical history can still be achieved.

A discreet user authentication model is introduced in Figure 2, where the interface of the authenticator can be a human officer and he is unnecessary to be a trusted party; authentication is mainly done by a feature matching module (usually as a secured software system). The module is responsible for generating less-sensitive questions based on the supplementary information from the attributes of a medical condition. Upon receiving the answers to those questions, the module then deduces a hypothetical answer; this hypothetical answer will be cross-checked with the actual answer that is read directly from the microchip of the smart card over a secure smart card reading channel. For an example if the user is suffering from hypothyroidism and this particular illness is being used for authentication, the feature matching module first gathers a list of less-sensitive questions from a mass database (that represents and generalizes the illness) such as what the average basal metabolic rate is, the intake of seafood, and experiences of any twitch in muscle. The questions are based on symptoms of a disease which are relatively less embarrassing to be communicated in an open channel. The answers will then be used to infer or predict a hypothetical disease. After the secure module reads the actual answer from the smart card owned by the user, an attempt of matching the hypothetical disease to the actual disease indicates whether the testing user who offered the answers is the authentic user.

By this design the secrecy which is the illness records of the user stored in the smart card will never leave the authentication system and hence will not be revealed directly to the public. The user will not be questioned directly about the illness (the secrecy), instead by asked by questions about his general lifestyle, dietary habits, and disease symptoms which he experienced. Based on this information, a hypothetical illness is inferred automatically inside the authenticator model which is processed by secure computer software. The human officer needs not to know anything about the user's medical history except to convey the questions that are generated by the system to the user and to input the user's answers back to the system for analysis.

The general workflow of the proposed privacy preserving authentication model is summarized by the following steps.

*Step 1.* Preparing knowledge models for each disease based on the mass medical dataset.

*Step 2.* When a user is presented for authentication, his card is first read and one of the illnesses is randomly selected for testing.

*Step 3.* If no knowledge model exists for any of his illnesses, abort.

*Step 4.* From the knowledge model of the selected illness, derive a short list of questions about the symptoms and/or the lifestyle habitant attributes that lead to the illness (e.g., smoking habits lead to lung cancer). More details will follow in the next section of the paper.

*Step 5.* Signal the questions to the authenticator interface which is a human officer in this case.

*Step 6.* The questions are being asked from the user by the officer, the officer collects the answers.

*Step 7.* The answers are entered to the system, quantified, and processed.

*Step 8.* A hypothetical illness is estimated based on the answers, with a probability of likelihood.

*Step 9.* The system reads the user's biometric card via a secure smart card reader for the information about the illness IFF it has not been done so in Step 2. Otherwise, skip this step.

*Step 10.* Positive verdict is generated if the hypothetical and actual illnesses do match. Otherwise go to Step 12.

*Step 11.* Human officer is acknowledged about the result and decide accordingly to grant the authentication to the user.

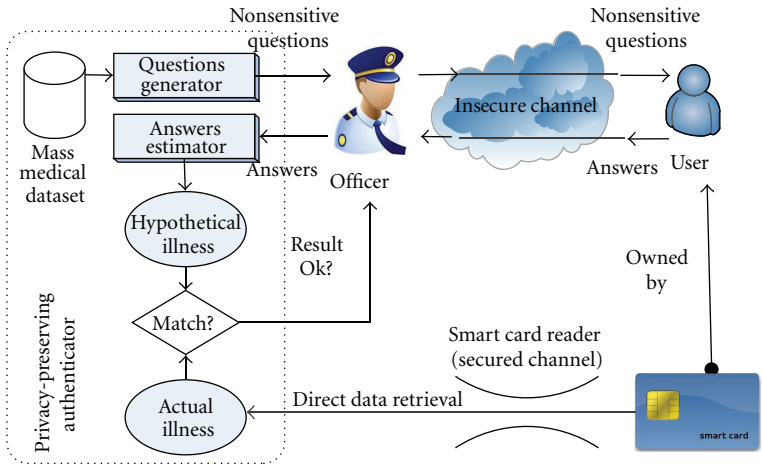



Figure 2: Privacy-preserving authentication model using medical history from biometric card.

*Step 12.* Case is rejected or is repeated from Step 2 by choosing another illness.

## 2. Design of the Privacy-Preserving Authentication Model

The prime challenge to be overcome by our proposed model is preserving the user's privacy by selecting a short list of useful features which are extracted from the medical data for use in authentication. The features which are being used instead of the direct information about the illness must satisfy two conditions: first, they should contain less sensitive elements and only a few of them should be used; using too many or a complete set of features will ultimately reveal the identification of the illness. Second, not only the features must be controlled in quantity but also they must be strongly relevant to the target illness such that the illness can be sufficiently characterized by only a handful of these features.

The principle for the protection of privacy to work is founded on causality which is defined as the relationship between an event (the cause) and a second event (the effect), where the second event is a consequence of the first. The term "feature" which we use here is the direct factor which is a factor that affects an effect directly, that is, without any intervening factors. For instance, lung cancer is due to smoking habit. The feature or direct factor in this example is smoking, and the effect as the consequence is lung cancer as the illness. The true identification of an illness is described by only a number of significant features. The features are allowed to be queried and responded in open, and the answers (values) to the features could effectively refer back to the same illness.

The design of the model which is shown in Figure 3 combines the three analytic approaches for supporting defining causality relations of medical attributes from some given clinical history data. The data are collectively accumulated from a sizeable population as reference, which is called mass medical data. The mass medical data are fed to a sequence of processes to generate five main types of information for quantitatively describing the causality among the features and the illness. We call this causality information which comes in four types: (1) correlation counts. The counts represent the linear relationships for each pair of features including the features to the class illness. (2) The optimal number of features that can be used to describe an illness. (3) The significance value of each feature; nonlinear relationships are inferred by decision tree induction which results in dependency network that shows the factors and their significances pertaining to the outcome of a disease, and a set of decision rules that represent the nonlinear and sometimes even complex relationships of the factors. (4) the relation strength between each pair of features. (5) The cooccurrences of the features with values that describe a state of the illness. Nonlinear relationships refer to a varying trend that describes the outcome, often by more than one factor. These five types of causality information would be used along with the new input testing values of the features (resulted from asking the user the questions) to estimate a hypothetic illness by summarization in data mining. If the hypothetic illness is the same as the target illness, the feature values which are provided from the authentication questions would also be the same as the feature values derived from the target illness. Our model should be extensive enough to cover the attributes' relations/dependencies both linear and nonlinear and by finding such relations quantified for matching for authentication.

*2.1. Building Knowledge Models.* Knowledge models must initially be built prior to authentication application. Mass medical history data collected from the public consists of many patients' past records, each record spanning across a collection of attributes, that are to be used in building decision trees for finding the underlying relations. Each record often includes attributes taken from measurements of tests, diagnosis and demographic attributes of the patients' profiles. The records from the mass database should be in the same format as the medical history database embedded in the card of the testing user. Our proposed model has a work flow that accepts medical history datasets that are structured

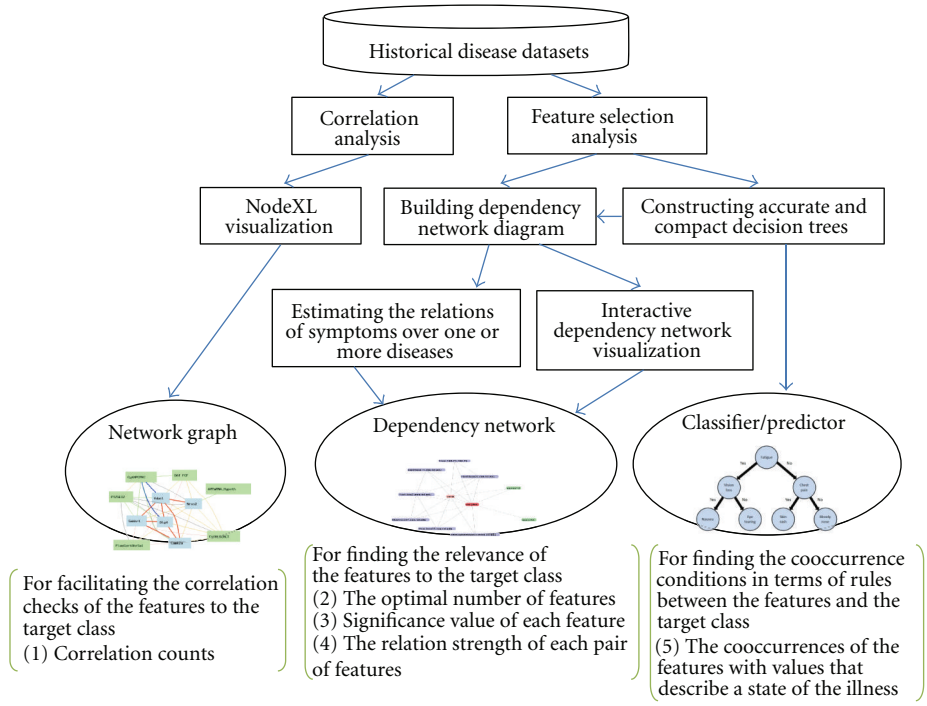



Figure 3: Model for deriving the attribute relations via Network Graph, Dependency Network and Rule-based Decision Tree.

in columns as attributes and rows as patients' records; computational processes that transform them into resultant outputs. There are mainly three streams of processing: (1) correlation analysis, with the aid of NodeXL visualization for generating Network Graph; (2) feature selection analysis, building an accurate and compact rule-based decision tree, extracting rules that show conditional relations among the attributes; (3) after feature selection analysis, merge multiple disease datasets, construct a dependency network and compute relation strengths among the attributes. From top to bottom, the original historical datasets are transformed through a sequence of subtasks which are described in details in the following sections. In our experiment here we verified this model by using two datasets lung cancer and heart disease, acquired from UCI Dataset Repository which is well known for benchmarking machine-learning techniques in computer science research community [5].

Pertaining to knowledge discovery in medical field, Ohsaki et al. compared the performance of 40 different interestingness measures via a rule-discovery experiment on clinical datasets of meningitis and hepatitis [6]. The results supported that a stable and reasonable performance is achieved by chi-square measure which is a prominent member of the family of information gain methods. This encourages us to follow along this direction for deriving useful rules for representing the relations between attributes and the class illness. Applying information theoretic techniques has its edge over frequency or statistical due to the nature of the data; linear trend implies a direct relation between a pair of univariate attributes. For multivariate attributes which are usually the case for high-dimensional medical data, the relations are cross-dependent among the attributes. Some recent work applied computational intelligence techniques that include Artificial Neural Network combined with Rough Set Theory [7] for extracting decision rules from medical data, Classification Rules with aid of Concept Lattice [8] for analyzing medical diagnostic data. The learning techniques based on information theoretic have been proved their usefulness as a tool for drawing conclusions from medical data. Ohsaki et al. [6] expanded the work by considering that attributes of different significances may be conditional (interdependent) in data classification and decision making. An attribute that has low significance close to zero may get omitted in the feature selection process but this attribute when used together with others may consequently lead to an important rule that represents useful knowledge. Therefore experiments were conducted in [6] that proved: if an attribute group which contains significant attributes, the attribute group must be significant and if an attribute group includes attributes with low significance individually, the attribute group possibly may have high significance. Subsequently this proof advocates that taking a singular view on the significance of individual attributes is not enough. Conditional relations among those attributes regardless of their significances must be taken into account in analyzing medical data. Therefore the three levels of analysis were proposed in our model design that allows users to find linear and nonlinear relations among data via Network Graph and Dependency Network, respectively as well as a rule-based decision tree that extracts and exhibits conditional rules for studying the conditional relations among the attributes.

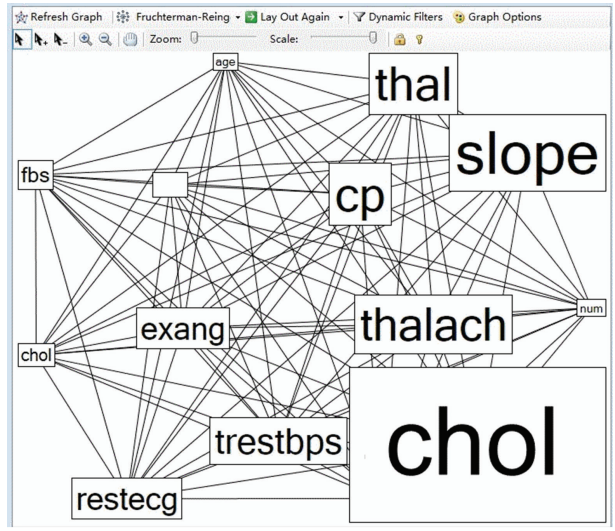
*2.2. Correlation Model.* By computing the correlation coefficients over the dataset, the strengths of the relations between

each pair of attributes could be obtained. Pearson algorithm is used as it is popular, simple, and powerful in evaluating the pairwise trend between two attributes, with value close to 1 means highly correlated. The purpose of finding a direct and linear relation between the attributes is twofold. First, medical professionals may be interested to know or to confirm which pairs of attributes are directly related for the sake of intellectual curiosity. For example, a person's weight and height are usually strongly correlated in the BMI calculation. There may exist some not-so-well-known kind of direct relationship in the process of knowledge discovery in different diseases or medical phenomena. Recent discoveries by correlation analysis include "High fizzy soft drink consumption linked to violence among teens (10/2011)", "TV Viewing Linked to Unhealthy Eating (09/2011)", and "Junk Food Makes Kids Fatter, But Happier (04/2009)" just to name a few. In our case of authentication, we want to first match the features of the test samples by the correlation values of their peer features. If two sets of features (test and reference) have a similar set of correlation values which are in two-dimensional form, the features are indeed similar and they are likely to infer to the same disease. From the performance view of authentication, this is a quick test that could be conducted first before proceeding to further complicated tests.

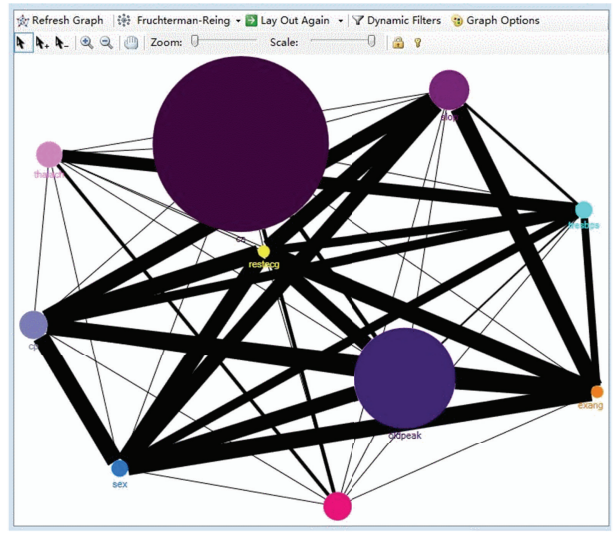
The second purpose is for finding redundant attributes and possibly eliminating them. Sometimes having fewer attributes among the various medical examination tests may be desirable in diagnosis of a disease. The authors in [8] applied context reduction technique to reduce those redundant attributes from the rules generated from classification tree. The motive for excluding redundant attributes in medical test is to replace expensive tests by cheaper tests (with less testing attributes).

Although correlation is a major criteria for manifesting similarities in medical analysis between data attributes [9], another criteria which may be equally if not more popular are association rules. In our model, the correlation coefficients matrix could be in turn filled in by a composite measure of support and confidence accordingly for association rules. Association rules take forms of $X \rightarrow Y$ where $X$ and $Y$ are sets of attributes. Support of $s$% means that $s$ number of records includes both $X$ and $Y$. Confidence of $c$% means $c$ amount of records that consists of $X$ must also contain $Y$. In our experiment, we computed correlation coefficient matrix as shown in Table 1.

Visualizing the correlation lattice as a Network Graph for the medical data is enabled by a software program called NodeXL-Network Overview, Discovery and Exploration for Excel, (software is freely downloadable from http://www.codeplex.com/). It is a free and open source spreadsheet add-in with features of network analysis and visualization. The information to be visualized is stored as a correlation coefficient matrix (Table 1) which is to be represented by a network graph. The attribute relations are represented as a column of graph edge information; they specify which pairs of vertices being connected in the graph network. In particular, the edges and vertices that are mapping the relations and attributes, respectively, have visual




(a)




(b)

Figure 4: Network Graph by NodeXL—Top: (a) attribute info. bottom: (b) relation strength.

properties to be programmed by the user according to the values in the correlation coefficient matrix, such as color, size, and shapes. In our case, only size is taken as a performance variable that represents the magnitude or strength, that is, strongly correlated relations between pairs of attributes take on thicker lines; attributes that occur more frequently in association rules are represented by bigger vertices.

Interactively, users can adjust settings of the control panel of the NodeXL template and explore the direct relations between attributes. We modified the visualization of edge thickness by using an exponential boosting function because the differences between the correlation coefficients are very small considering the value ranging from 0 to 1, often in decimal of one or two places. Screen captures of the Network Graph in NodeXL are shown in Figures 4(a) and 4(b) that show the distribution of frequently appeared attributes in

Table 1: Correlation coefficient matrix shows a segment of correlated attributes from our experimental data.

| | Item 1 | Item 2 | Item 3 | Item 4 | Item 5 | Item 6 | Item 7 | Item 8 | Item 9 | Item 10 | Item 11 | Item 12 | Item 13 | Item 14 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Item 1 | 1 | | | | | | | | | | | | | |
| Item 2 | **0.304628** | 1 | | | | | | | | | | | | |
| Item 3 | **0.086859** | **0.955484** | 1 | | | | | | | | | | | |
| Item 4 | **0.268803** | **−0.64095** | **−0.8371** | 1 | | | | | | | | | | |
| Item 5 | **−0.16671** | **0.876819** | **0.904209** | **−0.69252** | 1 | | | | | | | | | |
| Item 6 | **−0.83893** | **−0.50437** | **−0.42444** | **0.255924** | **−0.04098** | 1 | | | | | | | | |
| Item 7 | **−0.16783** | **0.880265** | **0.913895** | **−0.71313** | **0.999567** | **−0.055** | 1 | | | | | | | |
| Item 8 | **−0.76751** | **0.151117** | **0.226658** | **−0.21787** | **0.598202** | **0.775897** | **0.586164** | 1 | | | | | | |
| Item 9 | **−0.16533** | **0.88728** | **0.933745** | **−0.75615** | **0.995487** | **−0.08997** | **0.997845** | **0.554419** | 1 | | | | | |
| Item 10 | **−0.14374** | **0.896828** | **0.963708** | **−0.826** | **0.975886** | **−0.1679** | **0.981833** | **0.477619** | **0.992128** | 1 | | | | |
| Item 11 | −0.16987 | 0.881941 | 0.920922 | −0.7297 | 0.998566 | −0.06535 | 0.999708 | 0.576789 | 0.999127 | 0.986037 | 1 | | | |
| Item 12 | −0.84519 | −0.4055 | −0.32866 | 0.194064 | 0.06833 | 0.993808 | 0.053754 | 0.840339 | 0.017242 | −0.06468 | 0.042858 | 1 | | |
| Item 13 | −0.38102 | 0.759633 | 0.843887 | −0.74226 | 0.973958 | 0.126083 | 0.975193 | 0.712566 | 0.973904 | 0.956482 | 0.975849 | 0.229056 | 1 | |
| Item 14 | #DIV/0! | #DIV/0! | #DIV/0! | #DIV/0! | #DIV/0! | #DIV/0! | #DIV/0! | #DIV/0! | #DIV/0! | #DIV/0! | #DIV/0! | #DIV/0! | #DIV/0! | 1 |

association rules and correlation of attributes, respectively. For simplicity in the illustration, only the top 10 items (attributes) are shown in bold. The heart disease dataset originally has 76 attributes, describing the patients' health background, blood pulse rates and other measurements, and so forth.

*2.3. Feature Selection Analysis.* Along with the Correlation analysis which is an independent process by itself, our model suggests Feature Selection Analysis to be done before commencing to analysing nonlinear relations. Feature selection process has a long history in data mining whose aim is to selectively retain only the "useful" attributes, which are also known as features, in characterization of the data model prior to training a data mining model. In our case, feature selection allows us to compute a significance value for each feature, thereafter the selected features and their significance values will be used to construct a dependency network and a decision tree. For authentication, the matching will be done upon only a set of selected or "qualified" features that have high significance. By using a shorter list of important features, the time taken for the authentication process can be shortened.

A comprehensive survey on feature selection [10] describes many types of techniques for selecting useful attributes while filtering irrelevant ones. The technique that we adopted here is Information Gain that is shown to offer consistent performance from a collection of medical datasets from UCI. The characteristics of some widely used feature selection techniques are briefly listed in Table 1. What they have in common is the ability to evaluate the information entropy in such a way whether including the attribute under test would contribute to reducing the chaos of information or not. At the same time, this measure implies how much this particular attribute contributes to increasing the predictive power of the training model, therefore it is taken as a performance indicator for evaluating how a prediction outcome would depend on each attribute.

The method of using feature selection is slightly different in our model than in traditional data mining. Instead of directly short-listing top-$k$ worthy attributes to build a decision tree, in between we tried out all these algorithms and plot out three performance charts on worthiness of attribute, prediction accuracy, and decision tree size, by varying the number of the short-listed attributes who passed the feature selection test in ascending order. The attributes are first ranked and progressively one by one being added to the decision tree building process. The motive of this novel technique is to find a just enough amount of most highly contributing attributes. And also the attributes yield an optimal balance of accuracy and decision tree size.

From Figure 5, we can clearly see that an optimal number of qualified attributes to be used is 13, that is where the cross-point for the curves used by different feature selection algorithms. It is obvious that using too few would result in an inaccurate model, too many attributes mean expensive diagnosis tests. The number 13 which is deemed appropriate for including the most qualified attributes in training

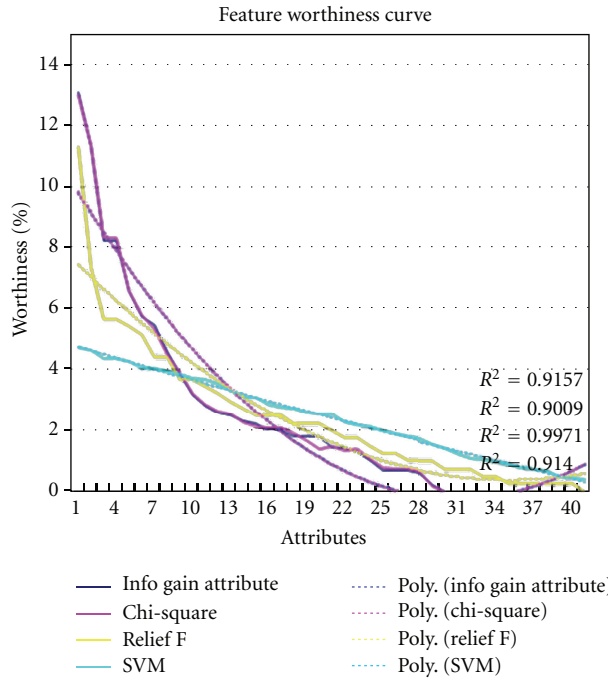



Figure 5: Worthiness measure of varying number of features being added into training a decision tree.

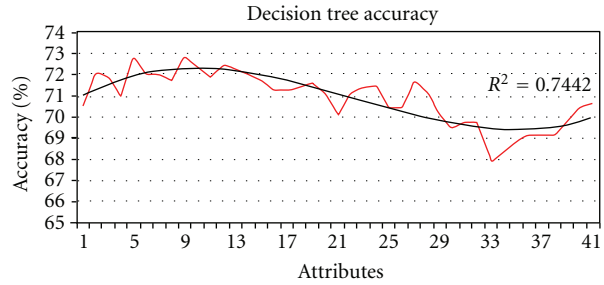



Figure 6: The accuracy of the decision tree model trained by using different number of qualified features.

a decision tree model is unanimously agreed to be the optimal point for accuracy (in Figure 6) and for tree size (in Figure 7) as well.

To recap, finding the significance values via Feature Selection analysis helps estimating the optimal number of most contributing features in building a decision tree and the significance values would be passed on to the next process, building a dependency network diagram. More importantly, from a handful of important features the authentication system can randomly set a subset from them for formulating questions every time.

*2.4. Dependency Model.* After obtaining a rule-based decision tree model, such as Ripper (Pruning to Produce Error Reduction) by William Cohen of AT and T Laboratories which is chosen in our model because of its suitability and relatively high accuracy, the information about the information gain for each attribute would be used for

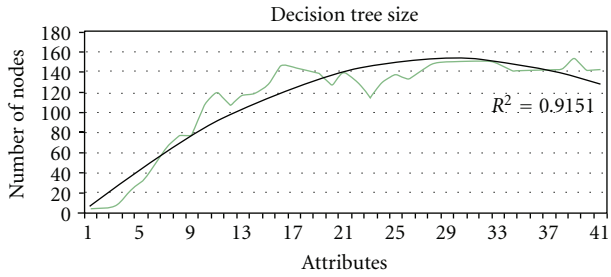



Figure 7: The resultant tree size of the decision tree model trained by using different number of qualified features.

inferring a collection of links each represents the predictive power (in term of information gain) towards the prediction class. Ripper has been shown performing fast with low error rate in accuracy [11]. For a simple illustrative example, in patients' records, we may find a very strong correlation between obesity and body weight, but they have no predictive power to diabetes disease on par with glucose level in blood. These attributes, however, may all pass the feature selection test as described earlier. We therefore opt to have a visual tool that interactively allows medical professionals to explore not only the direct (or linear) relations but predictive power which we loosely define it as "dependency" to a particular disease outcome.

Dependency Network Browser (DNB) is a standard data mining tool with Analysis Server by versions of Microsoft SQL 2000 and upwards. This tool is to present the dependencies or relationships among attributions in a data mining model. A decision tree would first be required to be built in order to display the predictive powers of the attributes in form of dependencies (arrows that connect from the attribute nodes to the prediction class). Once in the Dependency Network Browser, the trained decision tree model is expressed as a network of attribute nodes such that it offers the users the ability to view the whole prediction model from the perspective of all attributes by relationship information, therefore a global view of how attributes or factors contribute to prediction of a certain class.

In our experiment, some modifications were done on the standard copy of Dependency Network Browser, using Flash and NET programming framework. One major modification is to incorporate the ability of loading multiple medical history datasets so that dependencies can be traced across different diseases. This feature is useful for factors exploration especially those that were not previously known. The implicit link could be traced down a chain of diseases provided that they have common attributes in the forms of factors and symptoms in the sense of cause-and-effects (causality) by considering their relationships towards some related diseases. The second modification is a set of formula for quantitatively deriving a relational measure for this indirect dependency across diseases. The whole approach was coined as Extended Dependency Network Browser or eDNB for short.

One upfront technical challenge in implementing eDNB is the need of merging two or more medical datasets that have different dimensions in columns and rows though they may share some common attributes. This is known as schema matching and it is a classical problem in information integration. A number of automated methods have been attempted in the past [12], such as matching the missing values by textual similarity, guessing the figures by using the mean numbers, by most frequently appearing numbers, and so other statistical tricks on. For the demand of very high accuracy, however, in medical data analysis, we resort to the most accurate yet computational-intensive method by building a RIPPER decision tree for estimating the blank values. As long as the two medical datasets have sufficient amount of common attributes and the attributes have fairly good predictive powers to the diseases, the decision-tree-permissing-attribute method works satisfactorily. A pioneer work on applying decision trees for estimating missing values demonstrated its feasibility [13].

For our experiment, a number of decision tree types and methods have been attempted, like pure tree induction, rule-based methods, and Meta which means combined methods. It was found that RIPPER still offers the highest accuracy for our two datasets, with missing data estimation accuracy 69.2% and 69.9%, respectively, for heart disease and lung cancer. When the two datasets are successfully merged and the corresponding missing data are estimated, a rule-based decision tree model is generated, so is the eDNB. From the rule-based decision tree, RIPPER, some significant rules are extracted as examples below. The decision tree grows one rule at a time by adding antecedents to the rule until the rule reaches a perfect accuracy. The procedure searches for every possible value of each attribute and it selects the condition with highest information gain.

*Rule 1.* (sex >= 1) and (age >= 57) → class_heart_disease = 1 (79.0/23.0).

*Rule 2.* (sex >= 1) and (years <= 27) and (age >= 46) → class_heart_disease = 1 (46.0/19.0).

*Rule 3.* Otherwise → class_heart_disease = 0 (151.0/40.0).

In this example, it shows that the attributes gender, years of smoking, and age are conditionally related pertaining to predicting a disease outcome. Such relations do not show up in the Network Graph by measuring the correlation coefficient. But they have certain dependencies in the forms of information gain and predictive power towards a disease class. This is what our eDNB is supposed to essentially reveal.

In order to generalize our eDNB model in the methodology, algebraic equations are used to define the computation of relations between attributes. The diagram in Figure 8 shows a generic dependency model of two diseases $d_1$ and $d_2$, and the related factors or symptoms are associated with them as $s_0^{d_1}$ to $s_n^{d_1}$ for disease $d_1$ and $s_0^{d_2}$ to $s_n^{d_2}$ for disease $d_2$. Between the two diseases they possess common attributes such as $s_0^{d_1 \cup d_2}$ to $s_n^{d_1 \cup d_2}$, they are predicting both diseases.

Let $r_{di}(s_x, s_y)$ be the relation of a pair of symptoms which are predicting a common disease, $d_i$. For an example

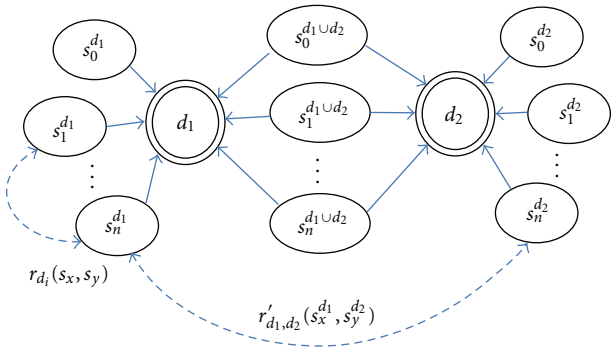



FIGURE 8: A network dependency model for two diseases and associated factors.

of disease $d_1$, the relation between symptoms $s_0$ and $s_2$ is expressed as:

$$r_{d_1}(s_0, s_2) = \frac{I_g^{d_1}\left(s_0^{d_1}\right) + I_g^{d_1}\left(s_2^{d_1}\right)}{\sum_{i=0}^{n} I_g^{d_1}\left(s_i^{d_1}\right)}, \tag{1}$$

where $n$ is the number of attributes predicting $d_1$. $I_g$ is the information gain value calculated in the feature selection process and decision-tree-building process for each attribution. Let $r'_{di,dj}(s_x^{di}, s_y^{dj})$ be the indirect relation of a pair of symptoms which are predicting two different diseases. For example, the indirect relation between $s_3$ from $d_1$ and $s_5$ from $d_2$ can be expressed as follow:

$$r'_{d_1,d_2}\left(s_3^{d_1}, s_5^{d_2}\right) = \begin{cases} 0 & \text{condition} \\ w_{d_1} \dfrac{I_g^{d_1}\left(s_3^{d_1}\right)}{\sum_{i=0}^{n_{d_1}} I_g^{d_1}\left(s_i^{d_1}\right)} & \\ \quad + w_{d_2} \dfrac{I_g^{d_2}\left(s_5^{d_2}\right)}{\sum_{i=0}^{n_{d_2}} I_g^{d_2}\left(s_i^{d_2}\right)} & \text{otherwise,} \end{cases} \tag{2}$$

where condition = if $d_1$ and $d_2$ have no predicting attribute in common, and $w_{d_1}$ and $w_{d_2}$ are weights proportional to the relative importance of the diseases that the common attributes are predicting about. The sum of the weights equals to 1. The weights are needed because the common attributes are assumed to be the only linkage between the two diseases and the relative portions of predictive powers by the common attributes deciding how important the disease is in relation to the symptoms.

$$w_{d_1} = \frac{\sum_{i=0}^{n_{d_1 \cup d_2}} I_g^{d1}\left(s_i^{d_1 \cup d_2}\right)}{\sum_{i=0}^{n_{d_1 \cup d_2}} I_g^{d1}\left(s_i^{d_1 \cup d_2}\right) + \sum_{i=0}^{n_{d_1 \cup d_2}} I_g^{d2}\left(s_i^{d_1 \cup d_2}\right)},$$
$$w_{d_2} = \frac{\sum_{i=0}^{n_{d_1 \cup d_2}} I_g^{d2}\left(s_i^{d_1 \cup d_2}\right)}{\sum_{i=0}^{n_{d_1 \cup d_2}} I_g^{d1}\left(s_i^{d_1 \cup d_2}\right) + \sum_{i=0}^{n_{d_1 \cup d_2}} I_g^{d2}\left(s_i^{d_1 \cup d_2}\right)}. \tag{3}$$

The current model can be extended to a chain of diseases that go beyond two adjacent diseases. So we let $r''_{di\cdots dk}$ be an indirect relation of a pair of symptoms which predict more than two diseases.

$$r''_{d\alpha\cdots d\Omega}\left(s_x^{d\alpha}, s_y^{d\Omega}\right) = \begin{cases} 0 & \text{condition} \\ \dfrac{\sum_{j=0}^{m(j:\alpha\cdots\Omega)} \text{Sig}_j}{\sum_{j=0}^{m(j:\alpha\cdots\Omega)} \sum_{i=0}^{n_j} I_g^{dj}\left(s_i^{dj}\right)} & \\ \quad \text{otherwise,} & \end{cases} \tag{4}$$

where condition = if the relation chain of any two disease is broken, that is, common symptoms of any two diseases along the chain are missing or do not have sufficient worthiness values. The chain is defined by the link of possessing common attributes from $d\alpha$ to $d\Omega$.

$$\text{Sig}_j = \begin{cases} w_{dj} \cdot I_g^{d\alpha}\left(s_x^{d\alpha}\right) & \text{if } j = \alpha, \\ w_{d_1} \cdot I_g^{d\Omega}\left(s_y^{d\Omega}\right) & \text{if } j = \Omega, \\ \text{Sig}_1 = Y_1, & \text{For } p > 1, \\ \text{Sig}_p = \beta \cdot Y_{p-1} + (1-\beta) \cdot \text{Sig}_{p-1} & \\ \qquad \text{if } j \in [\alpha + 1 \cdots \Omega - 1]. & \end{cases} \tag{5}$$

$\text{Sig}_p = \beta \cdot (Y_{p-1} + (1-\beta) \cdot Y_{p-2} + (1-\beta)^2 \cdot Y_{p-3} + \cdots + (1-\beta)^k \cdot Y_{p-(k+1)}) + (1-\beta)^{k+1} \cdot \text{Sig}_{p-(k+1)}$ for any suitable $k = 0, 1, 2, \ldots$ The weight of the general significance of the link between two diseases $Y_{p-i}$ is $\beta(1-\beta)^{i-1}$ where, the coefficient $\beta$ represents the degree of weighting decrease, a constant smoothing factor between 0 and 1. A higher $\beta$ discounts further linkages faster. $\text{Sig}_p$ is the value of significance at any position $p$ along the disease chain.

$Y_p$ is the linkage strength at any position $p$, along the chain. $Y_p$ is defined by the proportion of common symptoms and their predictability powers in disease $p$, in relation to disease $p-1$ and disease $p+1$.

$$Y_p = \frac{\sum_{i=0}^{n_{d_p}} \left\langle I_g^{d_p}\left(s_i^{d_p \cup d_{p+1}}\right) + I_g^{d_p}\left(s_i^{d_p \cup d_{p-1}}\right) \right\rangle}{\sum_{j=0}^{2} \sum_{i=0}^{n_{d(p+j-1)}} I_g^{d(p+j-1)}\left(s_i^{d(p+j-1)}\right)}. \tag{6}$$

For demonstration purposes, we used two medical datasets whose attributes and significances towards a disease are displayed in an eDNB. Figure 9 shows that all the selected attributes of the two diseases are fully displayed. A small panel at the bottom of eDNB allows user to choose two symptom attributes; then automatically the corresponding relation of the two symptoms are displayed.

Figures 10 and 11 demonstrate the operation of eDNB in a mode where only common attributes are displayed of the two diseases. There is a slide-bar by which a user can adjust the viewing by the strength of the dependencies. When the slide bar moves down the minimum requirement for dependencies strength increases, such that only the attributes that have strong dependencies would remain. Attributes of relatively weaker dependencies fade away. This way, the

relations of the attributes-to-attributes and dependencies of the attributes-to-diseases can be explored interactively.

With the model that derives information and relations about the features in place, the following operation in pseudocode describes about how feature matching is carried out.

*Step 1.* Questions are generated from a short list of features which have passed the feature selection process. If the list is large enough, a subset of features are being chosen randomly and used to generate the questions. Questions are conveyed over to the user via the authentication officer.

*Step 2.* The answers of the questions are returned from the user to the authenticator.

*Step 3.* Based on the answers which are the values for the short-listed features, the first matching test is over the correlation tables. Retain and score about the degree of matching of those feature values are in correlation with the target feature values from the known illness. If the scores are satisfactory over a minimum user-defined threshold, proceed to the next step, or else abort, return no match.

*Step 4.* From the dependency network graph, sum up the strength values (or significance values) as percentage scores towards predicting the target illness by the passed features. The higher the percentage scores are, the more matching the hypothetical illness that are being described by the testing features. Usually a user defined threshold is needed to decide if the matching is successful or otherwise. 50% is used here for relaxed matching; a very high minimum threshold can be set if strict matching is required. If the matching test fails, abort and acknowledge the user about the failure; he may choose to try the authentication again and the system will select another target illness for testing next time. If matching succeeds, proceed to next step.

*Step 5.* The user is successfully authenticated. The system returns a positive acknowledge to the user. Just as an option, the authentication can be tightened by choosing a series of illnesses for feature matching. Of course the list of questions will proportionally become longer, so is the feature matching time.

## 3. Conclusion

An emerging trend in biometrics is to tap on users' historical data. Medical history data is one option that can uniquely describe well of a user. However, one of the main challenges by using medical history for identity authentication is the possible leak of privacy if the medical history were to be directly questioned on. In this paper a novel model is proposed for preserving the privacy of medical history by implicitly questioning the users using the features of the illness instead of the illness itself. The features of the illness are subtle and appear to be quite general when viewed individually. It was found that when a set of features were used collectively together, they are sufficient to infer the

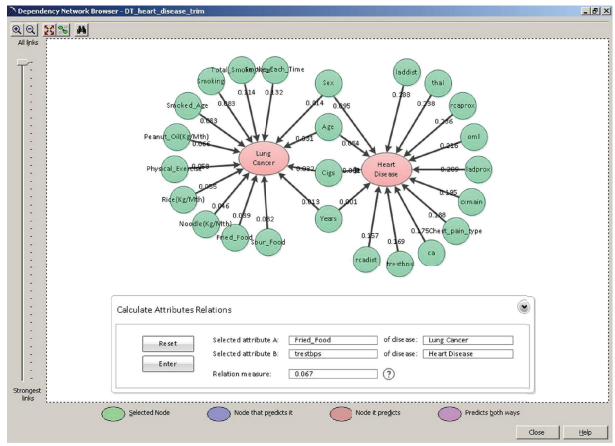



Figure 9: Visualising all attributes associated with the two diseases in eDNB.

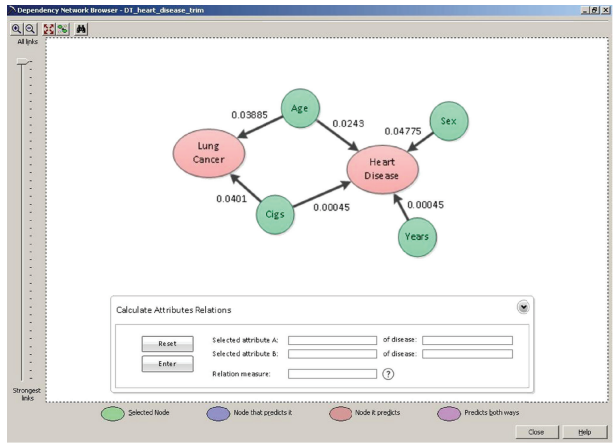



Figure 10: Visualising only the attributes which are in common between the two diseases in eDNB; the links that have relatively low predictive powers are dropped.

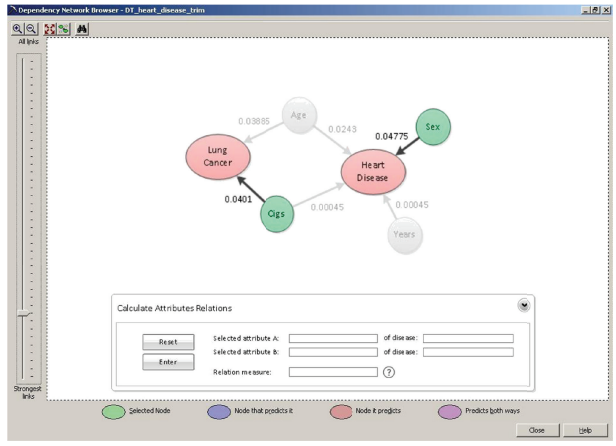



Figure 11: Visualising the common attributes that have relatively strong dependencies, others are grayed out.

identification of the illness. Taking the illness as the secrecy that is supposedly known only by the user, the questions that are derived from the selected features can be used to orally verify if the user knows of his past medical history: the experience of the illness indeed. From the answers of the questions that are derived from the selected features, a hypothetical illness is generated and it would be cross-verified by the illness data that was stored in a smart card. The emphasis of this authentication model is the causality that is the basis for quantifying relations between the features and the target illness. It is like a one-way hash that transforms a set of information into a target (illness) which we could use for matching it with the actual target stored on the card. A number of technical support functions are described in this paper; they are feature selection, correlation values computation, and dependency network. Though the foundation is laid by the contribution of this paper in preserving privacy in user authentication over medical history, a number of future works are possible. The matching process can be fine tuned by considering more than one illness, as the current limitation of the model is testing by one illness at a time. Some automated and intelligent process is needed to derive suitable questions from the selected features. And a performance evaluation should be conducted for checking the accuracy and speed of the whole authentication process too. These should be addressed in future works.

## References


[1] S. Mohamed, D. Noureddine, and G. Noureddine, "Face and speech based multi-modal biometric authentication," *International Journal of Advanced Science and Technology*, vol. 21, no. 6, pp. 41–56, 2010.

[2] A. Jagadeesan, T. Thillaikkarasi, and K. Duraiswamy, "Cryptographic key generation from multiple biometric modalities: fusing minutiae with iris feature," *International Journal of Computer Applications*, vol. 2, no. 6, pp. 16–26, 2010.

[3] N. Ann and T. Sotirios, "A study in authentication via electronic personal history questions," in *Proceedings of the 12th International Conference on Enterprise Information Systems (ICEIS '10)*, pp. 63–70, June 2010.

[4] M. Nishigaki and K. Makoto, "A user authentication based on personal history- a user authentication system using E-mail history," *The Journal on Systemics, Cybernetics and Informatics*, vol. 5, no. 2, pp. 18–23, 2007.

[5] C. J. Merz and P. Murthy, "UCI repository of machine learning database," ftp://ftp.ics.uci.edu/pub/machine-learning-databases/.

[6] M. Ohsaki, H. Abe, S. Tsumoto, H. Yokoi, and T. Yamaguchi, "Evaluation of rule interestingness measures in medical knowledge discovery in databases," *Artificial Intelligence in Medicine*, vol. 41, no. 3, pp. 177–196, 2007.

[7] A. Sakr and D. Mosa, "Dealing medical data with fundamentals of new artificial intelligence," *International Journal of Engineering Science and Technology*, vol. 2, no. 9, pp. 4406–4417, 2010.

[8] A. Gupta, N. Kumar, and V. Bhatnagar, "Analysis of medical data using data mining and formal concept analysis, world academy of science," *Engineering and Technology*, vol. 11, pp. 61–64, 2005.

[9] M. Strickert, F. M. Schleif, T. Villmann, and U. Seifferta, *Unleashing Pearson Correlation for Faithful Analysis of Biomedical Data, Similarity-Based Clustering*, Springer, Berlin, Germany, 2009.

[10] M. Dash and H. Liu, "Feature selection for classification," *Intelligent Data Analysis*, vol. 1, pp. 131–156, 1997.

[11] R. Abraham, J. B. Simha, and S. Iyengar, "Medical datamining with probabilistic classifiers," in *Proceedings of the 9th International Conference on Information Technology (ICIT '07)*, 2007.

[12] K. Jaewoo and J. F. Naughton, "Schema matching using interattribute dependencies," *IEEE Transactions on Knowledge and Data Engineering*, vol. 20, no. 10, Article ID 4527243, pp. 1393–1407, 2008.

[13] Y. Hang and S. Fong, "Aerial root classifiers for predicting missing values in data stream decision tree classification," in *Proceedings of the SIAM International Conference on Data Mining (SDM '11)*, vol. WS03, pp. 1–10, Mesa, April 2011.