Proposed Business Intelligence Models for Medical Risk Assessment

Case study of Venous Thrombosis Disease in Egypt

Dr. Edward Wadid[1], Dr. Nevine Makram Labib[2] and Prof. Sayed Abdel Wahab[1]

[1]Department of Computer and Information Systems
Faculty of Management Sciences,
Sadat Academy for Management Sciences
Cairo, Egypt
edwardwadid@gmail.com

[2]Department of Business Administration
Faculty of Business Administration, Economics and Political Science
The British University in Egypt, Egypt
nevmakram@gmail.com

Abstract— Risk assessment tools have been widely used in various fields such as Information Technology, Environmental studies as well as Healthcare. This paper explores the use of Business Intelligence tools in the healthcare industry in developing countries. In doing so, three different models using SQL Server 2008 Business Intelligence Tool were explored. These models are Naïve Bayes, Decision Trees and Neural Networks. Hence, a prototype Intelligent Risk Assessment Model, DVTRAM (Deep Vein Thrombosis Risk Assessment Model) is proposed. It applies different data mining techniques in order to uncover hidden patterns that may lead to medical complications such as Pulmonary Embolism (PE). Results showed that all of the three models were able to extract patterns in response to the predictable state. As for the performance of the models, they varied depending on the class value. In the future, the outcomes may constitute a good background for the development of a Medical Expert System in the domain of Internal Medicine.

Keywords- Business Intelligence (BI), Risk Assessment, , Data Mining (DM), Naïve Bayes, , Neural Networks, DVT, VTE

1. Introduction

Medical risk assessment has become a part of the daily activities of primary care physicians. It involves the identification of the risk factors, personal characteristics and test findings, which are associated with the increased incidence of a given disease, and the evaluation of the potential risk factors that may result out of it. The level of risk can be described either qualitatively (i.e. by classifying risk into categories as 'high', 'medium', or 'low') or quantitatively (with a numerical estimate).

The traditional risk assessment, using data analysis, has become insufficient, and methods for efficient computer-based analysis became essential. Examples of these methods are the Intelligent Data Analysis (IDA), Data Mining (DM) and Machine Learning.

As for Business intelligence (BI), it may be defined as "a set of mathematical models and analysis methodologies that systematically exploit the available data to retrieve information and knowledge useful in supporting complex decision-making processes"[1]. The BI tools are a type of application software designed to report, analyze and present the data previously stored in a data warehouse or data mart.

A BI system provides decision makers with information and knowledge extracted from data, through the application of mathematical models and algorithms. The rational approach typical of a BI analysis may be summarized in the following main characteristics. First, the objectives of the analysis are identified and the performance indicators that will be used to evaluate alternative options are defined. Then Mathematical models are developed by exploiting the relationships among system control variables, parameters and evaluation metrics and finally, what-if analyses are carried out to evaluate the effects on the performance determined by variations in the control variables and changes in the parameters. Some of the BI techniques are Data Mining (DM) that makes use of numerous methods for automatically searching large amounts of data for patterns and other interesting relations and Data Warehouses that use logical collections of information with structures that favor efficient data analysis (such as OLAP and Decision Support Systems (DSS) [2].

This research discusses the development of a risk assessment system using both Data Mining and Business Intelligence to support the specialists in defining the risk level of a certain disease. It investigates the potential of these data to predict the risk of a Venous Thrombosis (VTE) outcome for patients since an accurate risk prediction system may give clinicians an early indication of danger, thereby allowing enough time for medical

intervention or closer monitoring of the patient. While the medical aspect of this research is important, the central aim of this research is to present a practical approach and to investigate the exploitation of frequent patterns as an underlying technique for risk assessment purpose.
Hence, the goals of the research are to predict the risk level of DVT and to identify the significant influences and relationships in the medical inputs associated with the predictable state DVT.

2. Literature Review

As stated in one of the recent survey papers that dealt with the use of Data mining techniques in healthcare, for both the diagnosis and prognosis purposes, the following algorithms were found out to be of high performance: Decision Trees, Support Vector Machine, Artificial neural networks , Naïve Bayes and Fuzzy Rules. Analyses showed that it is very difficult to consider a single data mining algorithm as the most suitable for the diagnosis and/or prognosis of diseases since the performance of the algorithms depends mainly on the case as some of the cases require a combination of different algorithms in order to provide effective results [3].
Regarding the Deep Venous Thrombosis (DVT) disease, which is the main concern of this paper, a study made use of a genetic algorithm to construct decision trees model so as to predict the presence of the disease. It was found out that although the Decision trees are simple and practical as prediction models, they can be complex and incomprehensible [4].
Another study dealt with the task of predicting which patients are most at risk for post-hospitalization VTE, given a set of cases and controls. For this purpose, machine-learning methods were used to induce models for the prediction. Several risk factors for VTE that were not previously recognized were identified and the study showed that machine-learning methods were able to induce models that identify high-risk patients with accuracy that exceeds previously developed scoring models for VTE [5].
A third study investigated the DVT risk in patients with relapsed chronic lymphocytic leukemia treated with lenalidomide. It was found out that these data linked lenalidomide associated with DVTs with TNFα upregulation and endothelial cell dysfunction and suggested that aspirin may have a role for DVT prophylaxis in these patients [6].
A research reported an evaluation of a computerized tool to identify patients at high risk for VTE that found a sensitivity of 98% and positive predictive value of 99%. It also mentioned another computer program that was used to detect VTE and had a sensitivity of 92%, specificity of 99% and a positive predictive value of 97% to identify DVT and a sensitivity of 100%, specificity of 98% and positive predictive value of 89% to identify PE. It showed that these tools were found to provide a dependable method to identify patients at high risk for and with VTE [7].

3. Medical Problem of the Case Study

A deep-vein thrombus (blood clot) is an intravascular deposit that is composed of fibrin and red blood cells with a variable platelet and leukocyte component. Deep-vein thrombosis occurs when a thrombus forms (usually in regions of slow or disturbed blood flow) in one of the large veins, usually in the lower limbs, leading to either partially or completely blocked circulation.

A clot blocks blood circulation through these veins, which carry blood from the lower body back to the heart. The condition may result in health complications, such as fatal Pulmonary Embolism (PE) that can occur when a fragment of a blood clot breaks loose from the wall of the vein and migrates to the lungs, where it blocks a pulmonary artery or one of its branches. When that clot is large enough to completely block one or more vessels that supply the lungs with blood, it can result in sudden death. Deep Vein Thrombosis and PE are collectively known as Venous Thromboembolism (VTE). Since DVT has a high mortality rate, predicting it early is important [8].

4. Model Development Methodology

The proposed model, DVTRAM (Deep Venous Thrombosis Risk Assessment Model), uses the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology and the Data Mining Extensions (DMX), a SQL-style query language for data mining, for building and accessing contents of the models [9].

4.1. CRISP-DM Methodology

According to the CRISP DM Methodology, the DM process consists of three stages:

1. *Initial exploration:* that starts with the data preparation.
2. *Model building or pattern identification:* that involves considering various prediction models and choosing the best one based on their predictive performance.
3. *Deployment:* that involves using the model selected in the previous stage and applying it to new data in order to generate predictions and estimations of the expected outcomes.

4.2. Data Collection Methods

Two types of data collection methods were used. They are the following:

1. Literature review was conducted for the state of knowledge of risk factors of VTE.
2. Questionnaire: Based upon the evidence presented in the literature review and the experts' opinions, a

questionnaire was developed for medical specialists to collect their opinions concerning the estimation of risk levels for each risk factor. Another questionnaire was developed to collect patients' data, including risk factors, based upon the previous questionnaire and the experts' opinion. These data were the inputs of the mining models of the research.

Many problems have been faced, while trying to collect the needed data, such as the availability of medical data; as they were only available in a paper format since there were no medical records comprising such data.

The data were extracted from surveys taken from 6 Hospitals across Egypt and medical cases from some specialists of Hematology diseases. All data collected from hospitals conform to the patients' data privacy and security regulations. These data are considered de-identified. Identifiable means the data that is explicitly linked to a particular individual along with the data that include health information with data items that could reasonably be expected to allow individual identification. Hence, 600 patient cases have been collected in paper format then converted into digital format.

As for loading these data, Microsoft Excel Spreadsheets were used to enter data in a flat file as an initial phase then it was converted into a database using MS SQL 2008.

The database was then explored to be better acquainted before using these data in the core DM process. This exploration was done using simple SQL queries that consist of statistical analysis and aggregations, and graphical visualization.

4.3. Data Preparation Phase

This step was concerned about deciding which data will be used as input for DM methods in the subsequent step. Preparing data for the mining process consisted mainly of combining all of the relevant data in one table, or dataset, so that it acts as the source for the learning algorithms, and also dividing it properly between training and test sets. The training dataset was used to build several DM after being pre-analyzed so as to see how the attributes were represented in terms of their values in order to determine the initial input set of attributes.

5. Description Of Data

The database comprises the medical records of 408 patients (after being preprocessed) extracted from 6 hospitals. Each patient record includes a patient ID and a list of up to ten risk factors.

5.1. Initial Feature Selection

The analytical dataset is comprised of several attributes. However, some of them did not carry any relevant information from the analysis perspective. For instance, the attribute 'Long distance travel' for patient is missing as no data were available for this factor. In all of the cases there were no available data about genetic risk factors and other female risk factors such as pregnancy or hormone replacement therapy.

Table 1 reviews the attributes that have been selected for the analysis and those that have been rejected.

TABLE 1 Initial feature selection

Attribute	Accepted	Reason for Rejection
Gender	yes	
Age	yes	
BMI	yes	
Smoking	yes	
Immobility	yes	
Alcohol	No	No available data for such Attribute
Long distance travel	No	No available data for such Attribute
Medical illness	yes	
Minor Surgery	yes	
Major Surgery	Yes	
Family History	Yes	
Previous History	Yes	
Pregnancy	No	No available data for such Attribute
Oral contraceptives	No	No available data for such Attribute
Hormone replacement therapy	No	No available data for such Attribute

5.2. System Overview

Before explaining the individual components, a high-level preview of the entire DVTRAM framework is provided in Figure 1. Since the objective of the research was to develop a system that can help in estimating the risk levels of DVT, the following system components were used. They are illustrated in figure 2.

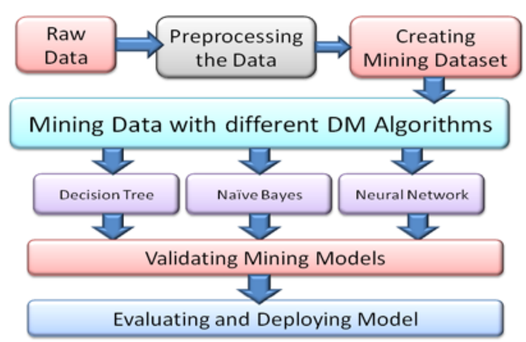

Figure 1 System Architecture

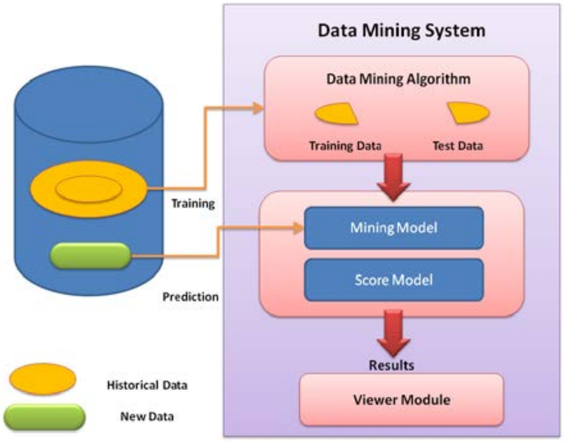

Figure 2 Components of Data Mining System

6. Mining Models with MS-SQL Business Intelligence

Microsoft SQL business intelligence tool has been selected for developing the different mining models for the proposed system.

6.1. Data Reception Phase (Analysis Module)

The records were split equally into two datasets: training dataset (204 records) and testing dataset (204 records). Records for each set were selected randomly to avoid bias. In this research, classification-modeling technique has been used as mining technique. The prediction model made use of three Data Mining model, Naïve Bayes, Decision Trees and Neural Networks. Naïve Bayes algorithm supports only categorical (discrete) attributes while Decision Trees and Neural network algorithms both support categorical and continuous attributes. To ensure consistency, categorical attributes have been used for all three models. We have identified the medical attribute "Risk Level" as the predictable attribute for patients risk level and the attribute "Patient-ID" was used as the key. All of the input attributes as explained in detail in table 2. As for data quality problems, such as noise and missing, inconsistent and duplicate data, they have been resolved in the datasets.

TABLE 2 Description of Attributes

S	Attribute name	Attribute Type	Attribute Value
1	Patient-Id	Key Attribute	Patient's identification number
2	Gender	Input Attribute	(value Male; value: Female)
3	Age	Input Attribute	Age in Year
4	BMI	Input Attribute	BIM in numbers
5	Smoking	Input Attribute	(value: Yes; value No)
6	Immobility	Input Attribute	
7	Hypertension	Input Attribute	(value Yes; value No)
8	Medical Illness	Input Attribute	Name of the medical illness
9	Minor surgery	Input Attribute	Name of the minor surgery
10	Major surgery	Input Attribute	Name of the major surgery
11	Family History	Input Attribute	(value Yes; value No, value don't know)
12	Previous History	Input Attribute	(value Yes; value 2 No)
13	Overall Risk	Predictable Attribute	Very low ,Low , Moderate , High, Very High

The trained model was evaluated against the testing dataset for their accuracy and effectiveness before they were deployed in DVTRAM.

The two methods used for evaluating the mining models were the Classification Matrix, which is a matrix for each model that specifies the Input Selection; it can quickly see how often the model predicted accurately, and the Lift Chart which compares the accuracy of the predictions of each model, and can be configured to show accuracy for predictions in general, or for predictions of specific value. Following is the evaluation of each model.

6.2. Naives Bayes Model

The Microsoft Naive Bayes does not introduce any specific constraints other than for the numbers of attributes. These numbers are limited with the use of the model's parameters. Also the method requires the input attributes to be discrete. The model of the Naive Bayes was built with the default setting of the parameters. The exception is the "Minimum Dependency Probability = 0.005". Tests have shown that the outcome of the method was affected by the modification of the parameters, because they mostly concern the number of attributes and their states. Results are summarized in table 3.

Table 3 Classification Matrix by Percentages for Naive Bayes model

	High (Actual)	Low (Actual)	Moderate (Actual)	Very High(Actual)	Very Low(Actual)
High	73.97 %	0.00 %	20.55 %	41.03 %	0.00 %
Low	0.00 %	81.82 %	1.37 %	0.00 %	0.00 %
Moderate	10.96 %	0.00 %	76.71 %	0.00 %	0.00 %
Very High	15.07 %	0.00 %	0.00 %	58.97 %	0.00 %
Very Low	0.00 %	18.18 %	1.37 %	0.00 %	100.00 %
Correct	73.97 %	81.82 %	76.71 %	58.97 %	100.00 %
Misclassified	26.03 %	18.18 %	23.29 %	41.03 %	0.00 %

Figure 3 represents the accuracy chart for Naïve Bayes model. The blue line represents the 'no model', the red line is 'the ideal model' and the green line represent the

Naïve Bayes model. From the graph it could be seen that the Naïve Bayes model is quite near the ideal model.

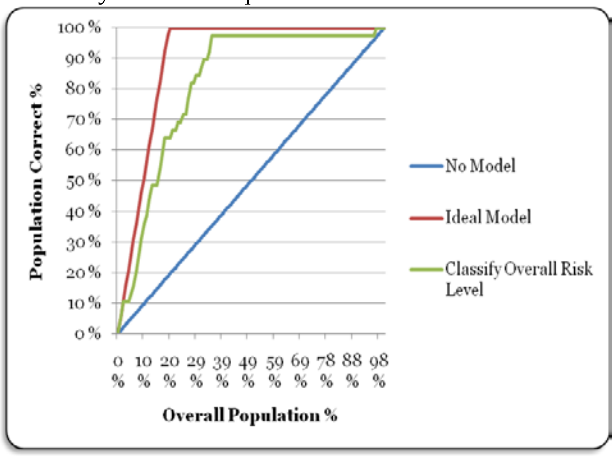

Figure 3 Accuracy Chart for Naive Bayes Model

6.3. Decision Tree Model

The Microsoft Decision Tree model incorporates features of the C4.5 and the CART algorithms. Thus, they are capable of performing predictions both in discrete and continuous problems. A tree can be grown on training data which contains errors. The algorithm does not implement pruning. Instead, the growth of a tree is controlled in two ways: Bayesian score – a score which stops further growth of a tree if the remaining data does not justify any more splits and Parameter COMPLEXITY_PENALTY – a parameter which takes values from 0 to 1, where the higher the value the smaller the tree as illustrated in table 4.

Table 4 Classification Matrix by Percentages for Decision Tree Model

	High(Actual)	Low(Actual)	Moderate(Actual)	Very High(Actual)	Very Low(Actual)
High	77.61 %	0.00 %	17.65 %	27.91 %	0.00 %
Low	0.00 %	85.71 %	11.76 %	0.00 %	31.58 %
Moderate	7.46 %	14.29 %	70.59 %	0.00 %	68.42 %
Very High	14.93 %	0.00 %	0.00 %	72.09 %	0.00 %
Very Low	0.00 %	0.00 %	0.00 %	0.00 %	0.00 %
Correct	77.61 %	85.71 %	70.59 %	72.09 %	0.00 %
Misclassified	22.39 %	14.29 %	29.41 %	27.91	100.0

6.4. Neural Network Model

The Microsoft Neural Network is an implementation of the feed-forward neural network (no cycles in the graph are allowed). There are two types of functions associated with each neuron: combination and activation. Following are the results of using Neural Network model.

TABLE 5 Classification Matrix by Percentages for Neural Network model

	High(Actual)	Low(Actual)	Moderate(Actual)	Very High(Actual)	Very Low(Actual)
High	70.15 %	0.00 %	17.33 %	29.27 %	0.00 %
Low	13.43 %	88.89 %	24.00 %	0.00 %	16.67 %
Moderate	2.99 %	0.00 %	53.33 %	2.44 %	33.33 %
Very High	10.45 %	0.00 %	1.33 %	65.85 %	0.00 %
Very Low	2.99 %	11.11 %	4.00 %	2.44 %	50.00 %
Correct	70.15 %	88.89 %	53.33 %	65.85 %	50.00 %
Misclassified	29.85 %	11.11 %	46.67 %	34.15 %	50.00 %

7. Results and Medical Assessment:

7.1. Models Validation

As mentioned before, the Microsoft SQL Server implements only two performance measure techniques: a Lift Chart and Classification Matrix techniques. The X-axis shows the percentage of the test dataset that is used to compare the predictions. The Y-axis shows the percentage of values predicted to the specified state. The blue and green lines show the random-guess and ideal models respectively. The purple, yellow and red lines show the Neural Network, Naïve Bayes and Decision Tree models respectively. The top line (red) shows the ideal model; it captures 100% of the target population for patients with DVT using 50% of the testing dataset. The bottom line (blue) shows the random line which is always a 45-degree line across the chart. It indicates that if we are to randomly guess the result for each case, 50% of the target population would be captured using 50% of the testing dataset. All three model lines (purple, green and Light-blue) fall between the random and ideal lines.

The following figures show that all three models had sufficient information to learn patterns in response to the predictable state. Figure 4 illustrates the lift chart validation for High risk level patients.

All of three models were able to extract patterns in response to the predictable state (High). The most effective model to predict patients who are likely to have a defined risk level for DVT disease appears to be Naïve Bayes followed by Decision Trees and Neural Networks. Figure 5 illustrates the lift chart validation for Low risk level patients. Also all of three models were able to extract patterns in response to the predictable state (low). The most effective model to predict patients who are likely to have a defined risk level for DVT disease appears to be Naïve Bayes followed by Neural Networks and finally Decision Trees.

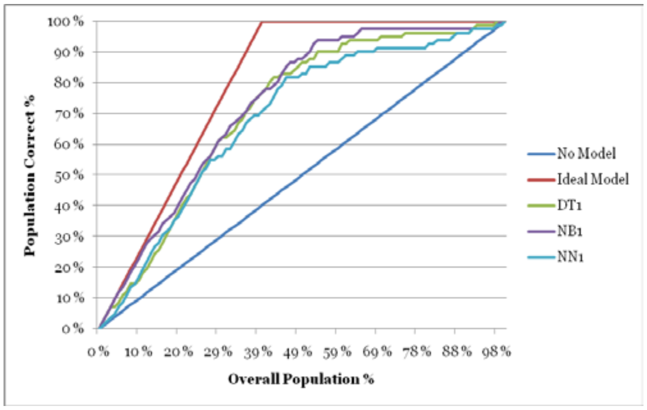

Figure 4 Lift Chart for High risk level patients

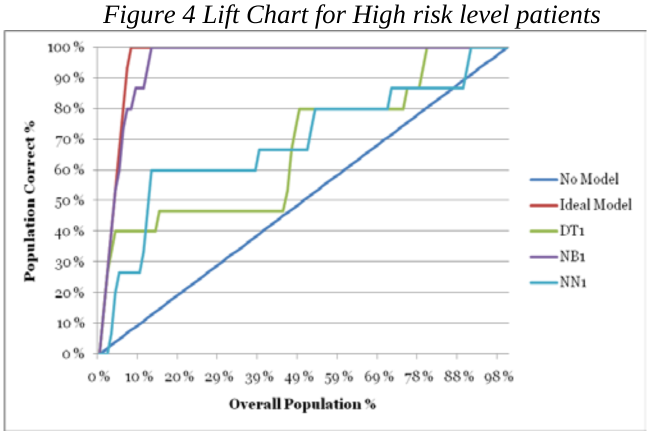

Figure 5 Lift Chart for low level risk

All three models achieved the objectives of the mining goals as they could provide good decision support to healthcare practitioners in assisting physicians and patients and discovering the medical factors associated with DVT disease.

7.2. Sample Case:

Hence, DVTRAM was able to support prediction queries based on "what if" scenarios. Users input values of medical attributes to diagnose patients with DVT disease. For example, entering the following attributes:

Gender = Male, Age = 71, BMI = 32, Smoking = Yes, Immobility = Use aid, Medical illness = Cancer, Minor Surgery = No, Major Surgery =No, Family History = No and Previous History = No into the models, would produce the results shown in Figure 6.

Figure 6 Result of DVTRAM system for a given data.

The three models ranked the person risk level within two risk levels. Naïve Bayes gave the Very High risk with probability (63%), the Decision Tree ranked in a High risk level with (43%) and Neural Network ranked in a High risk level with (64%). Based on these high figures, medical doctors can recommend that the patient is ranked between the high and very high risk level of DVT. Performing "what if" scenarios could thus help prevent a potential DVT occurrence.

7.3. Medical Assessment of the Results:

The previously mentioned results were revised by two Hematology specialists. They found them acceptable although they had some comments such as that the factors related to female gender they were concerned about did not appear in the assessment. In addition there was a clear confusion in the classification between Low and very Low risk levels and between High and Very High risk levels too. Some of the factors taken in consideration, such as major surgery and medical illness did not reflect the actual reality. As for Genetic characteristics, although they were the most important variables that determine the level of risk, they were summarized in a one factor, family history, which was not enough to clarify the relationship of different genetic factors with the disease. Therefore, this risk assessment system may be used as a kind of initial assessment only and specialists should be referred to in order to diagnose the situation carefully.

7.4. System Evaluation:

The mining goals, previously mentioned, were evaluated against the three-trained models.

Concerning the first goal, all three models were able to predict the risk level of DVT given patients' medical profiles using the singleton query and batch or prediction join query. As for the second goal, the system was able to identify the significant influences and relationships in the medical inputs associated with the predictable state DVT. The Dependency viewer in Decision Trees and Naïve Bayes models showed the results from the most significant to the least significant medical predictors. The most significant factor is Age followed by Medical Illness. Decision tree model gave a significant relation to all input attributes while Naïve Bayes gave a low significance to BMI attribute.

8. Summary and Conclusions

A prototype DVT disease risk assessment system was developed using three Data Mining classification-modeling techniques. DMX query language and functions were used to build and access the models. The models were trained and validated against a testing dataset. Accuracy Chart and Classification Matrix methods were used to evaluate the effectiveness of the models.

8.1. Contribution of the Research:

The research offers a contribution to the field of Business Intelligence and Medical risk assessment since the proposed system provides a Data Mining Tool for classifying patient risk characteristics based on features extracted from their medical data and acts as an intelligent system for estimating the risk level of suspected DVT patients. Eventually, these information will help specialists to use their resources more effectively.

8.2. Problems Faced :

They were primarily concerned with the data collection as the data were unreliable and difficult to extract. In some cases, the noise present in the samples was very high. As for the number of samples, it was not adequate to train the different models properly.

8.3. Limitations of the Research:

Following are some limitations of the work presented in this research paper:

1. The current version of DVTRAM is based on thirteen attributes. The list needs to be expanded to provide a more comprehensive diagnostic system.
2. It only used categorical data while for some diagnostic cases, the use of continuous data may be necessary.

8.4. Benefits and Future work of the Research

The system may serve as a training tool to train nurses and medical students to estimate patients risk levels of DVT disease. It can also provide decision support to assist medical doctors to make better clinical decisions or at least provide a "second opinion." The web version of the system can be used to assist anyone to determine his risk level for developing DVT. As for future work, the following enhancements can be made:

1. DVTRAM can be further enhanced and expanded so as to incorporate other medical attributes.
2. It can also incorporate other data mining techniques. Continuous data can also be added.
3. Text mining can be integrated with Data Mining.
4. The risk assessment model may be applied on other medical conditions and diseases.
5. Using different mining tools to testing and validating results rather than the Microsoft Data Mining tools.

9. References

[1] Carlo Vercellis ,"Business Intelligence: Data Mining and Optimization for Decision Making", John Wiley and Sons Ltd. Publication, 2009.

[2] http://en.wikipedia.org/wiki/Business_intelligence

[3] Elma Kolçe (Çela) and Neki Frasheri, "A Literature Review of Data Mining Techniques used in Healthcare Databases", ICT Innovations 2012 Web Proceedings - Poster Session ISSN 1857-728.8

[4] Christopher Nwosisi, Sung-Hyuk Cha, Yoo Jung An, Charles C. Tappert, and Evan Lipsitz, "Predicting Deep Venous Thrombosis Using Binary Decision Trees", IACSIT International Journal of Engineering and Technology, Vol. 3, No. 5, October 2011.

[5] Emily Kawaler, Alexander Cobian, Peggy Peissig, Deanna Cross, Steve Yale, and Mark Craven, "Learning to Predict Post-Hospitalization VTE Risk from EHR Data", AMIA Annual Symposium Proceedings. 2012; 436–445.

[6] Georg Aue, Jay Nelson Lozier, Xin Tian, Ann Marie Cullinane, Susan Soto, Leigh Samsel, Philip McCoy, and Adrian Wiestner, "Inflammation, TNFα, and endothelial dysfunction link lenalidomide to venous thrombosis in chronic lymphocytic leukemia", American Journal of Hematology, 86(10): 835–840, October 2011.

[7] R. Scott Evans, James F. Lloyd, Valerie T. Aston, Scott C. Woller, Jacob, S. Tripp, C. Greg Elliot and Scott M. Stevens, "Computer Surveillance of Patients at High Risk for and with Venous Thromboembolism AMIA Annual Symposium Proceedings 2010; 217–221.

[8] "Deep-Vein Thrombosis: Advancing Awareness To Protect Patient Lives" , White Paper Public Health Leadership Conference On Deep-Vein Thrombosis Washington, D.C., 2011.

[9] S. Palaniappan and R. Awang, "Intelligent Heart Disease Prediction System Using Data Mining Techniques", IJCSNS International Journal of Computer Science and Network Security, Vol. 8, No.8, August 2008, p. 343.