Predicting Deep Venous Thrombosis Using Binary Decision Trees

Christopher Nwosisi, *Member, IACSIT & IEEE*, Sung-Hyuk Cha, Yoo Jung An, Charles C. Tappert, and Evan Lipsitz

***Abstract*—An intrinsic disease where blood clots form in a deep vein in the body is known as *Deep Venous Thrombosis* (*DVT*). Since DVT has a high mortality rate, predicting it early is important. *Decision trees* are simple and practical prediction models but often suffer from excessive complexity and can even be incomprehensible. Here a *genetic algorithm* is used to construct decision trees of increased accuracy and efficiency compared to those constructed by the conventional ID3 or C4.5 decision tree building algorithms. Experimental results on two DVT datasets are presented and discussed.**

***Index Terms*-Decision Trees, DVT and Genetic Algorithm.**

I. Introduction

Deep Venous Thrombosis (DVT) is the formation of a *thrombus* (blood clot) in a deep vein in the body, typically in a lower leg or thigh. Figure 1 (a) shows a DVT patient with a painful and swollen leg. DVT is a disease of intrinsic origin and of great threat because it can occur without symptoms and have a high mortality rate [1]. While good medical technologies, such as *Duplex Ultrasonography* examination shown in Fig. 1 (b), can diagnosis DVT accurately, over two million DVT patients do not benefit from such technologies [2] .

To reduce the risk and to ensure that more patients can be treated before complications such as *pulmonary embolism* (PE) occurs predicting a DVT based on simple symptoms and medical history becomes critical. The prediction model must be as simple as possible.

Decision trees approximate discrete-valued target functions as trees and are widely used practical methods for inductive inference in knowledge discovery and decision support systems because of their natural and intuitive paradigm to classify a pattern through a sequence of questions [3]. Algorithms for constructing decision trees such as ID3 [3-5] and C4.5 [6] often use heuristics to find a shorter tree. Nevertheless, finding efficient and accurate decision trees is a difficult optimization problem [7, 8].

(a) Swollen leg due to DVT (b) DVT examination

Fig. 1 Swollen leg and duplex Ultrasonography.

TABLE I: DVT Dataset Attributes

	Name	Description
1	Sex (GN)	0 = female; 1 = male
2	Age (A6)	0 = age < 60; 1 = age ≥ 60
3	Diabetes (DB)	0 = normal; 1 = receiving some treatments
4	Smoking (SM) (SS, SB)	0 = never smoked; 1 = active Smoker; 2 = stopped smoking
5	Surgery (SR)	0 = never had surgery; 1 = previous surgery
6	Pain (PN) (LP, RP)	0 = none; 1 = pain in the leg {None, Right, Left, Bilateral}
7	Swelling (SW)	0 = none; 1 = swelling in the leg
8	Chest Pain (CP)	0 = none; 1 = pain in Chest
9	Cancer (CR)	0 = normal; 1 = positive
10	Cellulitis (CL)	0 = normal; 1 = positive
11	Injury (IJ)	0 =none; 1 = previous injuries
12	Pulmonary embolism (PE)	0 = never diagnosed; 1 = previously diagnosed
13	Congestive heart failure (HF)	0 = never diagnosed; 1 = previously diagnosed
14	Obesity (OB)	0 = none; 1 = specified
15	Accident (AC)	0 = none; 1 = had a fall
16	Hyperlipidemia (LIP)	0 = never diagnosed; 1 = previously diagnosed
17	Cardiac Dysrth-ythmia (CD)	0 = normal; 1 = previously diagnosed
18	Lymphoproliferat disease (LD)	0 = normal; 1 = previously diagnosed
	DVT	0 = negative for DVT; 1 = positive for DVT

This paper employs a method of constructing binary decision trees using a genetic algorithm as previously suggested [8]. Two data sets were extracted from the databases in the Montefiore Medical Center Vascular Laboratory and the general patient registry. Then, selected attributes were converted into binary attributes, and shorter and/or more accurate decision trees were created using the genetic algorithm on both of the DVT datasets.

Manuscript received April16, 2011, revised September 16, 2011.

C. Nwosisi is with the Computer Science Department, Pace University, White Plains, NY USA and Department of Thoracic and Cardiovascular Surgery, Montefiore Medical Center, Bronx, NY USA (e-mail: cnwosisi@montefiore.org).

S. Cha and C. C. Tappert are with the Computer Science Department, Pace University, White Plains, NY USA. (e-mail: scha@pace.edu; ctappert@pace.edu).

Y. An is with the Computer Science Department, Farleigh Dickenson University, NJ USA (e-mail: yoojung@fdu.edu).

E. Lipsitz is with the Department of Thoracic and Cardiovascular Surgery, Montefiore Medical Center, Bronx NY USA. (e-mail: elipsitz@montefiore.org).

The rest of the paper is organized as follows. Section 2 provides details of the DVT datasets, section 3 shows the decision tree experimental results on the two DVT datasets, and section 4 presents the conclusion and suggestions for future work.

II. DVT Binary Datasets

Known risk factors for *DVT* include diabetes, surgery, smoking, cancer, obesity, congestive heart failure, swelling, cellulitis, injury, and pulmonary embolism [9]. These factors can be determined by patients and physicians without medical examinations. Hence, eighteen potential attributes which can contribute to DVT were extracted from 515 records in databases at the Montefiore Medical Center Vascular Laboratory and the general patient registry. The dataset attributes are summarized in Table 1 together with the DVT outcome. Of the 515 records 350 patients were positive and 165 negative for DVT.

To use the genetic algorithm to build a binary decision tree, the attribute types must be binary [8]. The numeric data, 'age' attribute (A6) is binarized: 1 if over 60 and 0 otherwise. Non-binary nominal attributes include 'smoking' and 'pain' where they have three and four possible values, respectively. These are binarized as shown in Table 2.

TABLE 2: NOMINAL TO BINARY PREPROCESSING

Leg Pain			Smoking		
LP	RP		SB	SS	
1	1	Bi	1	1	Smoking
1	0	L	1	0	Stopped
0	1	R	0	0	Never
0	0	None			

The nominal type 'Leg Pain' attribute which has four possible values {L, R, Bi, N} in the original table is represented by two binary attributes, LP (pain in the left leg) and RP (pain in the right leg). The ternary attribute, 'Smoking' in the original table is represented by two binary attributes 'SB' (smoked before) and 'SS' (still smoking). Note that in certain datasets, the smoking attribute is denoted as simply 'SM' having either 0 (nonsmoker) or 1 (smoker). This is because not all questionnaires distinguish the stopped smoker. Similarly, the pain attribute may appear as simply 'PN' in some datasets.

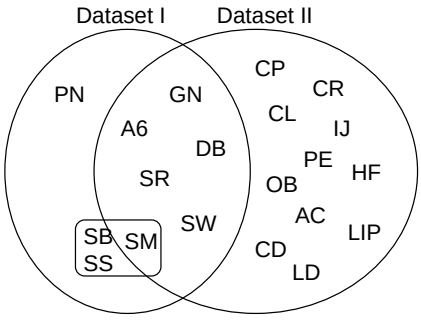


Fig. 2 Dataset I and II Relationship

In Fig. 2 we show the relationship between Dataset I and Dataset II. An overlap is depicted for the case when the tables share some common variables. The shared variables are gender, age, diabetes, surgery, swelling, and smoking.

Potential users for the proposed prediction models include patients at home and physicians. Two datasets were created – one for patients and one for physicians and those with medical knowledge. Because most patients have little medical knowledge, Dataset I was created with attributes which can be determined easily without much medical knowledge. Dataset II was created using all the attributes in Table 1 (except for PN) and this dataset is for physicians or users with some medical knowledge.

III. DVT Decision Trees

Consider the binary decision trees in Fig. 3 which are built from Dataset I. For each node the left branch is 0 (no) and the right branch is 1 (yes). Tree leaves indicate whether DVT is considered positive or negative.

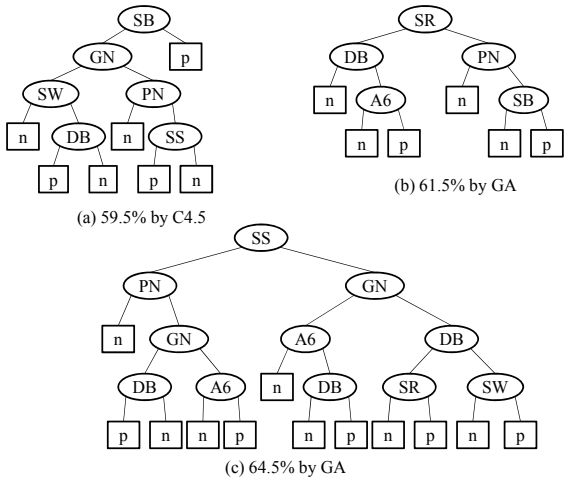


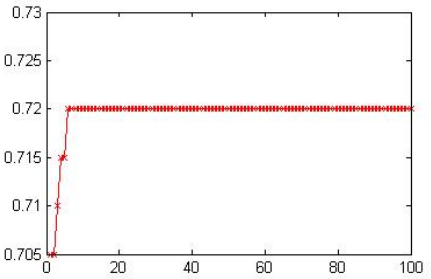
Fig. 3 Decision trees from Dataset I.

The decision tree in Fig 3 (b) suggests that a patient might have DVT if he/she never had surgery but has diabetes and is over 60 year old or might have DVT if he/she had previous surgery and feels pain in the leg and had previously smoked. The positive DVT cases can be logically expressed in the disjunctions of conjunctions form: $(SR = 0 \wedge DB = 1 \wedge A6 = 1) \vee (SR=1 \wedge PN=1 \wedge SB = 1)$.

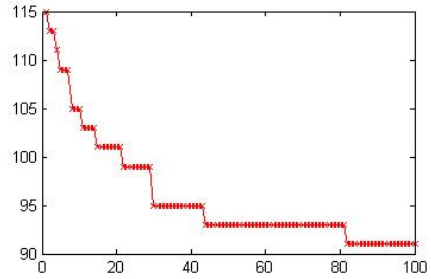
If a patient wants to predict the likelihood of DVT, the decision tree prediction model such as one in Fig. 3 (c) will prompt a sequence of questions. First, it will ask whether the patient is a current active smoker. When the patient answers with 'yes', it will prompt to ask about the gender. If the patient is a female, it will prompt whether she is over 60 year old. If the answer is "yes", it will ask whether she is a diabetic. If so, the decision tree predicts that she has a significant risk for DVT; in fact according to current laboratory records, one has a 66.67% chance of having a DVT under these conditions. Also, note that even though the decision tree predicts "No" in the left-most branch in Fig. 3 (c) where the patient is not currently smoking and does not feel pain, the chances that the patient may have DVT according to the database is about 45.6%. The decision tree is capable of providing the probabilities.

The popular decision tree algorithm C4.5 constructs pruned decision trees [6]; and was used to construct the tree shown in Figure 3 (a) having a performance of 59.5%. The

most basic and popular algorithm to construct decision trees, called *ID3,* constructs short trees [8]. However, the decision tree constructed by ID3 is not shown here because it was unreasonably large and too complex for patients and perhaps even physicians to use. However, its performance on Dataset I was 72% for DVT prediction.



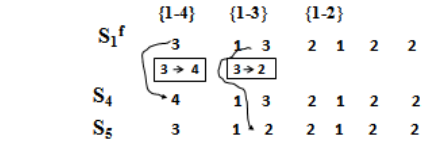
(a)

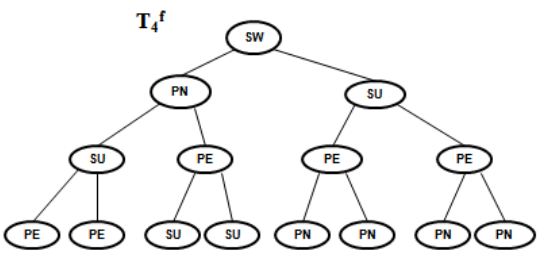


(b)

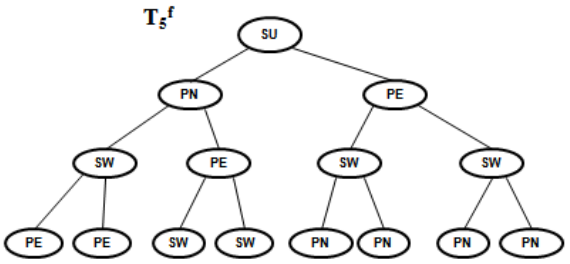
Fig. 4 Prediction rate (a) and number of questions (b) fitness functions of GA generations on Dataset I.

In this study, a genetic algorithm is used to find shorter and/or more accurate decision trees. It starts with 100 random decision trees, and only short and good decision trees survive to the next generation. Using mutation and cross-over operations, the next 100 generations are generated. Mutation and crossover are the two most common genetic operators. The mutation operator is defined as changing the value of a certain position in a string to one of the likely values in the range. Figures 5 illustrate the mutation process on the attribute selection scheduling string S_1^f = (3, 1, 3, 2, 1, 2, 2) and with P= (PN, PE, SU, SW). If a mutation occurs in the first position and changes the value to 4, which is in the range {1... 4}, T_4^f is generated. If a mutation happens in the third position and changes the value to 2, which is in the range {1… 3}, then T_5^f is generated. As long as the changed value is with the allowed range, the new string result will always generates a valid full binary decision tree.





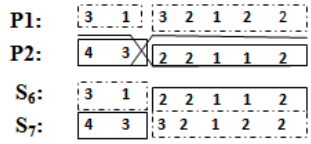
(a)

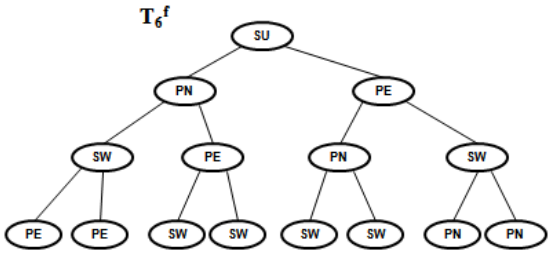


(b)

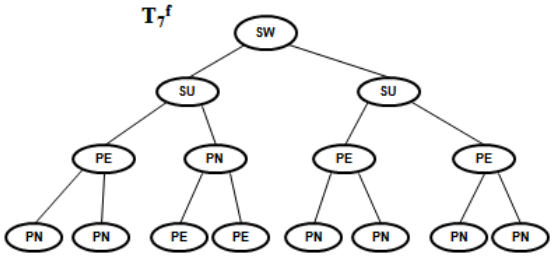
Fig. 5 Illustration of Mutation operator

Fig. 6 illustrates the crossover process by considering the two parents attribute selection scheduling strings, P1 and P2. After randomly selecting a split point, the first part of P1 and the last part of P2 contribute to yield a child strings S_6. Reversing the crossover produces a second child S_7. T_6^f and T_7^f full decision trees resulted from these two children. Fig. 4 (a) and (b) show the highest performance positive prediction rate and the lowest number of questions needed, respectively, to determine DVT for the entire test set for 100 generations.





(a)



(b)

Fig. 6 Illustration of Crossover operator

For dataset I, several decision trees which are shorter and more accurate than the one created by ID3 in Fig. 3 (a) were identified. A more accurate and shorter depth decision trees is shown in Fig. 3 (b) and an even more accurate one but of the same depth is shown in Fig. 3 (c).

For dataset II, Fig. 7 shows a decision tree by the C4.5 algorithm, and three decision trees by GA. The C4.5 decision tree is a skewed and deep (depth = 12) with an accuracy of 72.25%. When the tree is deep, strange rules can be found; for example, HF at the bottom of Fig. 7 (a) tree has the negative DVT when HF is positive, a rule which is not statistically valid.

To find shorter and more accurate trees, the GA was performed for 200 generations. By limiting tree depth to 5, the decision tree of Fig. 7 (b) was obtained. Its performance rate, however, is lower than that of C4.5. Fig. 7 (c) and (d) show trees found by limiting the tree depth to 6 and 7, respectively, and have accuracies of 73.75% and 75.25%. It has been observed that greater depth usually results in higher accurate until over-fitting occurs.

The best measure of efficiency (shortness) for a decision tree is probably the average number of questions required to obtain a prediction. Other measures might be the depth off the tree or the number of nodes in the tree.

the average number of questions increases monotonically with the depth limit, indicating that depth also appears to be a good measure of efficiency. The average number of questions to be asked of a user is 7.485 for the C4.5 decision tree in Fig. 7 (a) whereas there are several shorter ones listed in Table 3. The number of nodes is apparently not a good measure of efficiency – the C4.5 decision tree has 25 compared to 19, 32, and 44 in Fig. 7 (b), (c), and (d). From both a depth and average-number-of-questions perspective the complexity of the decision tree in Fig. 7 (d) can be considered much more efficient (simpler) than the decision tree from the C4.5 algorithm.

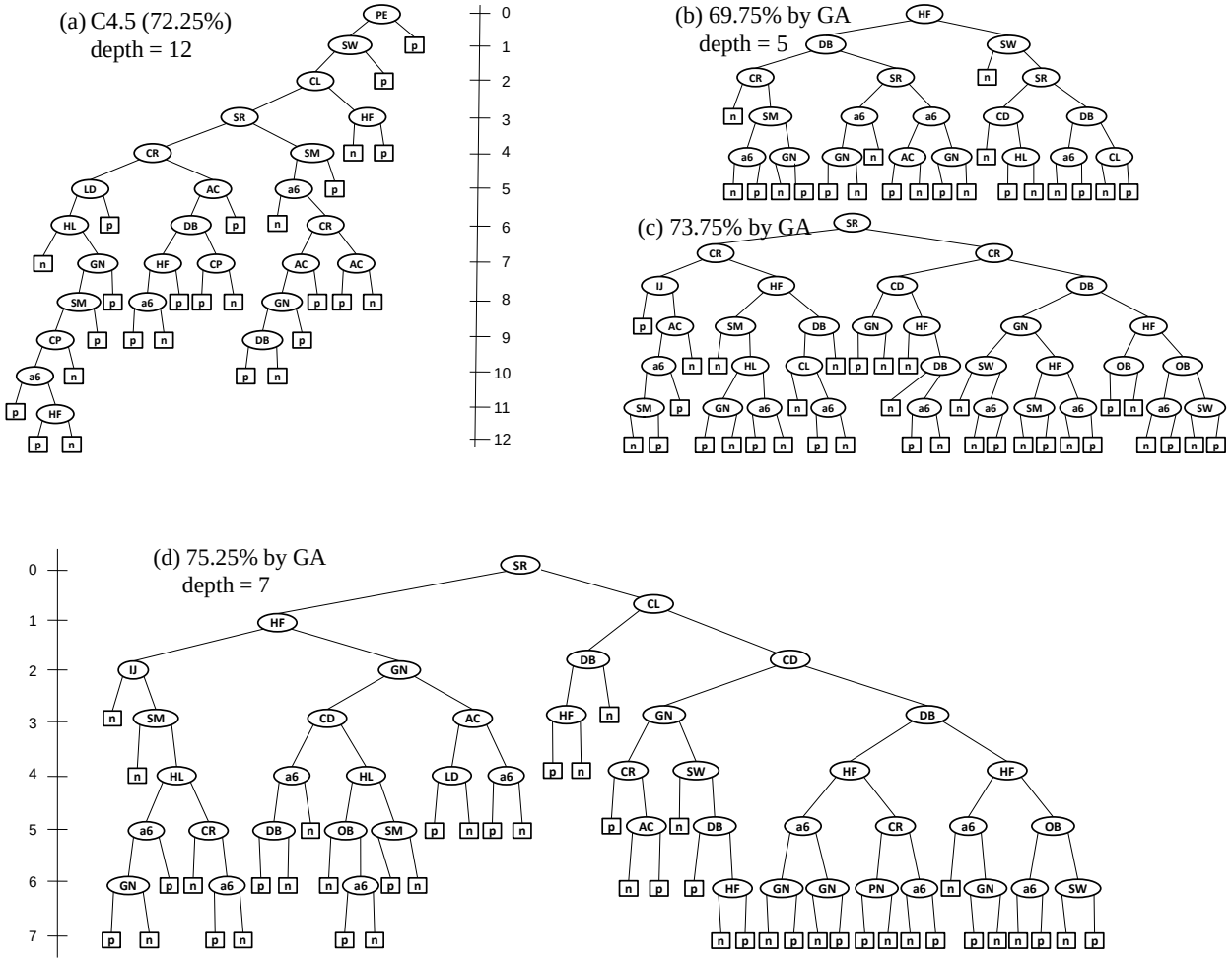


Fig. 7 Decision trees from Dataset II.

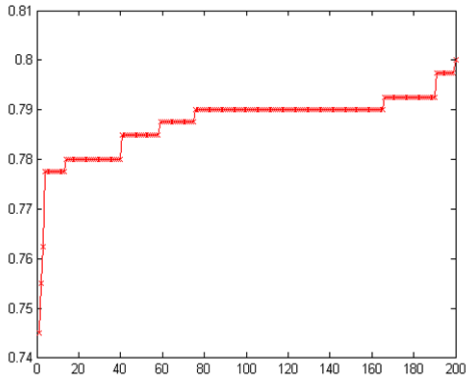
TABLE 3: COMPLEXITY OF DECISION TREES WITH DIFFERENT DEPTH LIMITS

Depth limit	Performance rate	The average # of question
5	69.75	2.9525
6	73.75	3.3725
7	75.25	3.8955
8	76.50	4.3275
9	76.75	4.8225
10	78.00	5.1225
11	78.50	5.4675
12	79.50	5.8675
13	80.25	6.3075

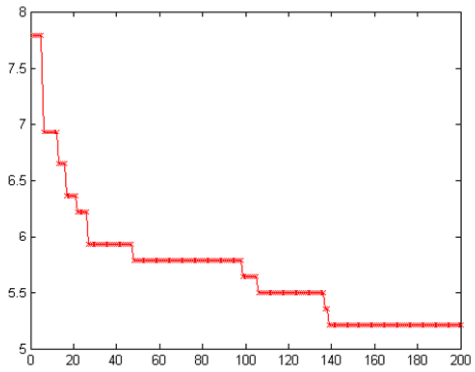
Table 3 shows the depth limits in GA, the performance rate, and the average number of questions to be asked. Note that

It was observed that accuracy increases as depth increases. At the depth of 12 the GA performance was 79.50 as compared with the C4.5 performance of 72.25 at the same depth. ID3 depth grows until the depth of 16 with a performance rate of 80% versus GA 80.25% with the depth of 13. These results clearly show that trees constructed by GA are both more accurate and more efficient.

Fig. 8 (a) and (b) show the highest performance positive prediction rate and the lowest number of questions needed, respectively, to determine DVT for the entire test set for 200 generations.



(a)



(b)

Fig. 8 Prediction rate (a) and number of questions (b) fitness function of GA generations on Dataset II.

IV. Discussion

For the purpose of DVT classification, the genetic algorithm is exploited to find shorter and/or more accurate decision trees than ones produced by the conventional ID3 and C4.5 algorithms. Experimental results on two datasets suggest that more accurate and efficient decision trees can be found by the GA. The efficiency (lower complexity) of a decision tree is best defined by the average number of questions asked to users, not by the number of nodes in the decision tree. In view of this argument, GA trees were found to produce more accurate and more efficient trees than ones produced by conventional methods such as the ID3 and C4.5 algorithms.

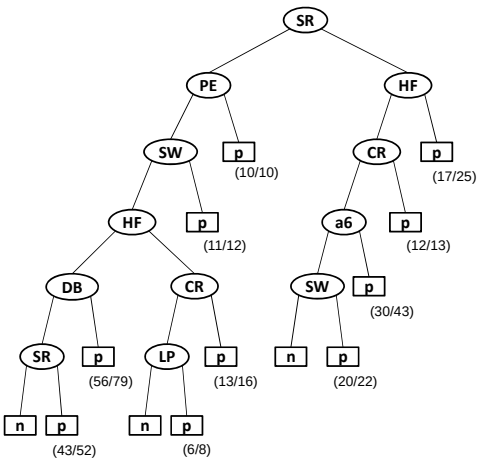


Fig. 9 DVT decision tree

The decision trees produced by the GA have significant clinical relevance. The results shown here increase the probability of predicting whether a patient would develop or have had DVT, which provides advancement in the diagnosis of DVT. The more efficient shorter trees add additional support for the GA method. Figure 9 shows a decision tree constructed with the input of experts after carefully reviewing the *forest* of good candidate decision trees found in this study. This might be the optimal decision tree based on the data and indicates that combining human knowledge and machine speed of processing can often produce a superior result than either the human or machine could produce separately.

With more iteration and deepening the depth of the tree, the decision trees produced by the GA depth limit clearly outperform the one produced by the ID3 method. This study introduced a simple decision tree to help lay people, medical technologists, and physicians identify the probability of a patient having DVT that prompts for testing before any complication occurs.

The decision trees found by using GA tend to be almost full binary trees, i.e., the width is large while the depth is short. For future work, the C4.5 pruning mechanism could be applied to decision trees produced by GA to make trees sparse and to further avoid the potential *over-fitting* problem.

Acknowledgment

The authors wish to thank Dr. Amit Shah, Dr. Hemal Shah and Josh Cruz for their support and for their useful comments.

References

[1] W. S. Moore, Vascular And Endovascular Surgery A Comprehensive Review. 7th ed. 2006: Saunders Elsevier.

[2] S. Z. Goldhaber, L. Visani, and M. D. Rosa, *Aute pulmonary embolism: Clinical outcomes in the International Cooperative Pulmonary Embolism Registry* (*ICOPPER*) Vol. 353. 1999: Lancet. 1386 - 19.

[3] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. 2nd ed. 2001: Wiley interscience.

[4] T. M. Mitchell, *Machine Learning*. 1997: McGraw-hill.

[5] J. R. Quinlan, *Induction of Decision Trees*. Machine Learning Research, 1986. 1(1): p. 81-106.

[6] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*. 2nd ed. 2006: Morgan Kaufmann.

[7] J. R. Quinlan, C4.5: *Programs for Machine Learning*. Morgan Kaufmann Publishers, 1993.

[8] S.-H. Cha and C. Tappert, *A Genetic Algorithm for Constructing Compact Binary Decision Trees*. Journal of Pattern Recognition Research, Vol 4 No 1, 2009, pp 1-13.

[9] Robert W. Zicker, *Deep Venous Thrombosis and Pulmonary Embolism in Bilateral Lower-Extremity Amputee Patients*. Academy of Physical Medicine Rehabilitation, 1999. 80: p. 509-511.

Dr. C. Nwosisi obtained his Doctorate Degree in Computing from Pace University, Master of Science in Management of Technology from Polytechnic University and BA in Computer Science from Hunter College of the City University of New York. He has co-authored several papers in Scientific Journals and International conferences.

Currently, he works for Montefiore Medical Center in the Bronx, New York, USA and as an Adjunct Professor at the College of Westchester in White Plains, New York, USA.

In 2010, he received the Upsilon Pi Epsilon Honors award for the Computing and Information Disciplines from Pace University. In 2007, he received the IEEE senior membership award. In 1993 and 1994 respectively, he was the recipient of the Recognition and Appreciation Awards from the Association for System Management. His current research interests include Machine Learning, Data mining and Pattern recognition.

Dr. Sung-Hyuk Cha received his Ph.D. in Computer Science from State University of New York at Buffalo in 2001 and B.S. and M.S. degrees in Computer Science from Rutgers, the State University of New Jersey in 1994 and 1996, respectively. From 1996 to 1998, he was working in the area of medical information systems such as PACS, teleradiology, and telemedicine at Information Technology R&D Center, Samsung SDS. During his PhD years, he was affiliated with the Center of Excellence for Document Analysis and Recognition (CEDAR). He has been a faculty member of Computer Science department at Pace University since 2001. His main interests include computer vision, data mining, pattern matching & recognition.

Dr. Yoo Jung An received M.S. and Ph.D. degrees in Computer Science from New Jersey Institute of Technology (NJIT) in January 2004 and 2008, respectively. During her graduate study years, she received a UPS Foundation Ph.D. Fellowship for Academic Excellence. Currently, she is a lecturer of the department of Information Systems and Decision Sciences at Fairleigh Dickinson University. Her professional activities include organizing the first ever workshop "The Semantic Web meets the Deep Web" as chair, at the IEEE Joint Conference on E-Commerce Technology and Enterprise Computing, E-Commerce and E-Services 2008, Washington D.C. Her research interests include Semantic Web, Health Informatics, Data Warehouses, Artificial Intelligence and Internet Security.

Dr. Tappert has a Ph.D. in Electrical Engineering from Cornell University and was a Fulbright Scholar. He worked on speech and handwriting recognition at IBM for 26 years, taught at the U.S. Military Academy at West Point for seven years, and has been a professor of computer science at Pace University since 2000. He has over 100 publications and his research interests include pattern recognition, biometrics, handwriting recognition/pen computing, speech recognition/voice applications, human-computer interaction, and artificial intelligence.

Dr. Evan Lipsitz obtained his M.D. degree from the College of Physicians and Surgeons of Columbia University. He completed residencies in General surgery at Columbia Presbyterian Medical Center and in Vascular Surgery at the Montefiore Medical Center and Albert Einstein College of Medicine.
Dr. Lipsitz currently serves as Chief of the Division of Vascular and Endovascular Surgery in the Department of Cardiovascular and Thoracic Surgery at the Montefiore Medical Center and Albert Einstein College of Medicine. He is also the Medical Director of the non-invasive Vascular diagnostic laboratory at the Montefiore Medical Center. He is a member of multiple professional societies and has published numerous authors in scientific journals.