Available online at [www.sciencedirect.com](http://www.sciencedirect.com/)



Expert Systems with Applications 34 (2008) 2465–2479

Expert Systems with Applications

[www.elsevier.com/locate/eswa](http://www.elsevier.com/locate/eswa)

AptaCDSS-E: A classiﬁer ensemble-based clinical decision support system for cardiovascular disease level prediction

Jae-Hong Eom a, Sung-Chun Kim b, Byoung-Tak Zhang a,\*

a *Biointelligence Laboratory, School of Computer Science and Engineering, Seoul National University, Seoul 151-744, Republic of Korea*

b *GenoProt Co. Ltd., 2FL Saeseoul Bldg., 94-1, Guro 6-dong, Guro-gu, Seoul 152-841, Republic of Korea*

Abstract

Conventional clinical decision support systems are generally based on a single classiﬁer or a simple combination of these models, showing moderate performance. In this paper, we propose a classiﬁer ensemble-based method for supporting the diagnosis of cardiovas- cular disease (CVD) based on aptamer chips. This AptaCDSS-E system overcomes conventional performance limitations by utilizing ensembles of diﬀerent classiﬁers. Recent surveys show that CVD is one of the leading causes of death and that signiﬁcant life savings

can be achieved if precise diagnosis can be made. For CVD diagnosis, our system combines a set of four diﬀerent classiﬁers with ensem- bles. Support vector machines and neural networks are adopted as base classiﬁers. Decision trees and Bayesian networks are also

adopted to augment the system. Four aptamer-based biochip data sets including CVD data containing 66 samples were used to train and test the system. Three other supplementary data sets are used to alleviate data insuﬃciency. We investigated the eﬀectiveness of the ensemble-based system with several diﬀerent aggregation approaches by comparing the results with single classiﬁer-based models. The prediction performance of the AptaCDSS-E system was assessed with a cross-validation test. The experimental results show that our system achieves high diagnosis accuracy (>94%) and comparably small prediction diﬀerence intervals (<6%), proving its usefulness in the clinical decision process of disease diagnosis. Additionally, 10 possible biomarkers are found for further investigation.

© 2007 Elsevier Ltd. All rights reserved.

*Keywords:* Clinical decision support system (CDSS); Cardiovascular disease; Classiﬁer ensemble; Support vector machines; Neural networks; Decision trees; Bayesian networks; Machine learning

1. Introduction
   1. *Background and motivation*

Recent surveys show that cardiovascular disease (CVD), which includes heart disease and stroke, is one of the lead- ing causes of death regardless of sex in the United States and all over the world ([CDC’s Report 1](#_bookmark22)). From the report, CVD accounts for nearly 40% of all deaths in the US annu- ally. While these largely preventable diseases are more pre- valent among people aged more than 65, the number of sudden deaths from heart disease among people aged 15–

34 has also increased substantially ([CDC’s Report 2](#_bookmark23)).

\* Corresponding author. Tel.: +82 2 880 1847; fax: +82 2 875 2240.

*E-mail address:* [btzhang@bi.snu.ac.kr](mailto:btzhang@bi.snu.ac.kr) (B.-T. Zhang).

Therefore, signiﬁcant life savings can be achieved if a pre- cise diagnosis can be made to CVD patients. Correct diag- nosis, however, is not easy to make and is often delayed due to the many factors complicating disease diagnosis. For example, clinical symptoms, functional, and patho- logic manifestations of heart disease are often associated with many other human organs besides the heart itself, and often heart disease may show diverse syndromes. Fur- thermore, diﬀerent types of heart disease can have similar symptoms, further complicating diagnosis ([Yan, Jiang,](#_bookmark58) [Zheng, Peng, & Li, 2006](#_bookmark58)).

To reduce the time of intensive diagnosis and to improve diagnosis accuracy, the development of reliable and power- ful clinical decision support systems (CDSSs) that support the aforementioned increasingly complicated diagnosis decision processes in the medical diagnosis is crucial ([Yan](#_bookmark58)

0957-4174/$ - see front matter © 2007 Elsevier Ltd. All rights reserved. doi:10.1016/j.eswa.2007.04.015

[et al., 2006](#_bookmark58)). Recently, many medical institutions are increasingly adopting tools that oﬀer decision support to improve patient outcomes and reduce clinical diagnosis errors and costs.

* 1. *Related work*

In the last two decades, the use of artiﬁcial intelligence tools has become widely accepted in medical applications to support patient diagnosis more eﬀectively. Especially, the application of various machine learning approaches such as decision trees (DTs), artiﬁcial neural networks (ANNs), Bayesian networks (BNs), and support vector machines (SVMs) have been actively tried for meeting clinical support requirements. Consequently, CDSS or medical diagnosis systems using diﬀerent machine learning approaches have shown great potential, and many machine learning methods have been tried for a wide variety of clin- ical and medical applications. Here we brieﬂy review some part of the previous work in this area before presenting our own machine-learning-based approach.

The use of *decision trees* is one of the most popularly applied methods for CDSS due to its simplicity and capac- ity for humanly understandable inductive rules. Many researchers have employed DT to resolve various biological problems, including diagnostic error analysis ([Murphy,](#_bookmark57) [2001](#_bookmark57)), potential biomarker ﬁnding ([Qu et al., 2002; Won](#_bookmark35) [et al., 2003](#_bookmark35)), and proteomic mass spectra classiﬁcation ([Geurts et al., 2005](#_bookmark36)).

*Bayesian networks* are a probability-based inference model, increasingly used in the medical domain as a method of knowledge representation for reasoning under uncer- tainty for a wide range of applications, including disease diagnosis ([Balla, Iansek, & Elstein, 1985](#_bookmark20)), genetic counsel- ing ([Harris, 1990](#_bookmark36)), expert system development ([Stockwell,](#_bookmark43) [1993](#_bookmark43)), gene network modeling ([Liu, Sung, & Mittal,](#_bookmark45) [2006](#_bookmark45)), and emergency medical decision support system (MDSS) design ([Sadeghi, Barzi, Sadeghi, & King, 2006](#_bookmark35)).

*Neural networks* have also been applied to the medical and diagnosis ﬁelds, most actively as the basis of a soft computing method to render the complex and fuzzy cogni- tive process of diagnosis. Many applications, for example, have shown the suitability of neural networks in CDSS design and other biomedical application, including diagno- sis of myocardial infarction ([Baxt, 1990, 1995](#_bookmark24)), diﬀerentia- tion of assorted pathological data ([Dybowski & Gant,](#_bookmark32) [1995](#_bookmark32)), MDSS for leukemia management ([Chae, Park,](#_bookmark25) [Park, & Bae, 1998](#_bookmark25)) and surgical decision support ([Li,](#_bookmark44) [Liu, Chiu, & Jian, 2000](#_bookmark44)), MDSS for cancer detection ([West](#_bookmark55) [& West, 2000](#_bookmark55)), assessment of chest-pain patients ([Ellenius](#_bookmark33) [& Groth, 2000](#_bookmark33)), decision making for birth mode ([MacDo-](#_bookmark47) [well et al., 2001](#_bookmark47)), heart disease diagnosis ([Tu¨ rkoglu,](#_bookmark48) [Arslan, & Ilkay, 2002](#_bookmark48)), CDSS for pharmaceutical applica- tions ([Mendyk & Jachowicz, 2005](#_bookmark54)), CDSS development for gynecological diagnosis ([Mangalampalli, Mangalampalli,](#_bookmark51) [Chakravarthy, & Jain, 2006](#_bookmark51)), and biological signal classiﬁ- cation ([Gu¨ ven & Kara, 2006](#_bookmark36)). Recently, multilayer percep-

trons (MLP), one of the most popular ANN models, has been applied to build an MDSS for ﬁve diﬀerent heart diseases diagnoses ([Yan et al., 2006](#_bookmark58)). The three-layered MLP with 40 categorical input variables and modiﬁed learning method achieved a diagnosis accuracy of over 90%.

*Support vector machines* are a new and promising classi- ﬁcation and regression technique proposed by Vapnik and his co-workers ([Cortes & Vapnik, 1995; Vapnik, 1995](#_bookmark34)). SVMs, developed in statistical learning theory, are recently of increasing interest to biomedical researchers. They are not only theoretically well-founded, but are also superior in practical applications. For medical, clinical decision sup- port and biological domains, SVMs have been successfully applied to a wide variety of application domains, including MDSS for the diagnosis of tuberculosis infection ([Veropo-](#_bookmark52) [ulos, Cristianini, & Campbell, 1999](#_bookmark52)), tumor classiﬁcation ([Schubert, Mu¨ ller, Fritz, Lichter, & Eils, 2003](#_bookmark35)), myocardial infarction detection ([Conforti & Guido, 2005](#_bookmark29)), biomarker discovery ([Prados et al., 2004](#_bookmark35)), and cancer diagnosis ([Majumder, Ghosh, & Gupta, 2005](#_bookmark49)).

*Hybrid models*. Besides single model-based approaches, hybrid machine learning approaches have also been tried to boost the performance of conventional single model methods and to overcome the inherent weaknesses in any single method. Many hybrid model approaches have been proposed, including a hybrid expert system for epileptic cri- sis decision using an ANN and a fuzzy method ([Brasil, de](#_bookmark26) [Azevedo, & Barreto, 2001](#_bookmark26)), an ANN with a DT for the development of an intelligent decision support system ([Tung, Huang, Chen, & Shih, 2005](#_bookmark46)), and an SVM with an ANN for electromyogram classiﬁcation ([Gu¨ ler &](#_bookmark36) [Koc¸er, 2005](#_bookmark36)). Recently, a novel SVM method in combina- tion with DT to generate human-understandable rules was proposed to alleviate the diﬃculty of understanding that arises from the black box characteristic of SVMs in trans- membrane segments prediction ([He, Hu, Harrison, Tai, &](#_bookmark36) [Pan, 2006](#_bookmark36)). Their approach achieved prediction accuracy of 93% with understandable prediction rules and with con- ﬁdence values over 90%.

*Ensemble models.* To overcome the limited generaliza- tion performance of single models and simple model com- bination approaches, more precise model combination methods, called ‘‘ensemble methods’’, have been suggested. This multiple classiﬁer combination is a technique that combines the decisions of diﬀerent classiﬁers that are trained to solve the same problem but make diﬀerent errors. Ensembles can reduce the variance of estimation errors and improve the overall classiﬁcation accuracy. Many ensemble-based approaches have been proposed in recent research, including an ANN ensemble for decision support system ([Ohlsson, 2004](#_bookmark59)), an ensemble of ANNs for breast cancer and liver disorder prediction ([Yang &](#_bookmark61) [Browne, 2004](#_bookmark61)), MDSS with an ensemble of several diﬀerent classiﬁers for breast diagnosis ([West, Mangiameli, Rampal,](#_bookmark53) [& West, 2005](#_bookmark53)), and multiple classiﬁer combinations with an evolutionary approach ([Kim, Min, & Han, 2006](#_bookmark37)).

* 1. *Objective and scope of the present work*

The majority of conventional CDSSs for disease diagno- sis are generally based on the symptoms of the patient or data from simple medical questionnaires. To our knowl- edge, a CDSS for CVD diagnosis using an ensemble of multiple classiﬁers for comprehensive diagnosis and possi- ble biomarker mining does not currently exist. The aim of this project is to develop a CDSS utilizing the expression information of physiological functional proteins with clas- siﬁer ensembles for patient diagnosis. The patient’s serum microarray chip data are analyzed with several diﬀerent classiﬁers in the ensemble. The developed system, Apta- CDSS-E (Aptamer biochip-based CDSS – ensemble ver- sion), supports physicians by providing supplementary diagnosis information and clinicians by providing a possi- ble set of biomarker candidates which can be used eﬀec- tively for practical CVD diagnosis after some further experimental veriﬁcations.

The rest of the paper is organized as follows: In Section [2](#_bookmark0) we outline the system architecture, describe several key components of the system, and review the four basis classi- ﬁers used in our proposed system for disease level classiﬁ- cation. In Section [3](#_bookmark8), the framework for constructing classiﬁer ensembles is presented. Experimental results are reported in Section [4](#_bookmark12), including data description, prepro- cessing and feature selection, quality analysis of data, the possible marker proteins discovered by the system, and dis- cussions of the results. Section [5](#_bookmark30) draws conclusions from this study.

1. The system architecture of AptaCDSS-E

The reviews of CDSS in literature show that very few studies involve ﬁeld tests of a CDSS and almost none use a naturalistic design in routine clinical settings with real patients. Moreover, the studies mostly concern physicians rather than other clinicians ([Kaplan, 2001](#_bookmark38)). On this point, in the development of AptaCDSS-E we considered both clinicians and physicians equally by providing diagnosis support information to physicians and by providing the information about possible biomarker candidates of dis- ease diagnosis to clinicians. The system can be used for CVD diagnosis in various ways such as a supplementary system for a periodic medical checkup or as a component of a hospital information system.

In AptaCDSSS-E, the patient diagnosis process starts from the doctor’s medical examination of a new patient by collecting blood samples when they need these blood analysis processes. Then, an aptamer biochip is created with the serum separated from the patient blood and pro- tein expression levels are scanned. Next, a new work list is created by the scanner interface and analyzed by the deci- sion engine of AptaCDSS-E trained with prior sample sets. The system provides integrated analysis results to the phy- sician, including clinical analysis facts. After the physi- cian’s ﬁnal decisions for a new patient, decision results are saved into the system database as a feedback informa- tion for future model updates and reﬁnements.

The system was implemented on the Microsoft Windows platform and has four major components. [Fig. 1](#_bookmark1) shows the

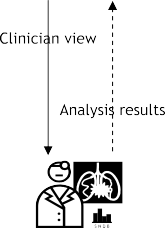
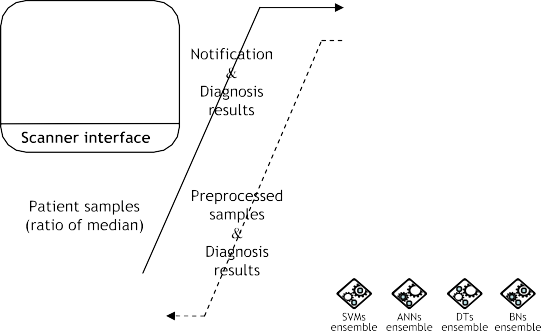
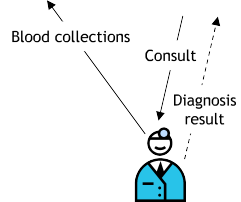
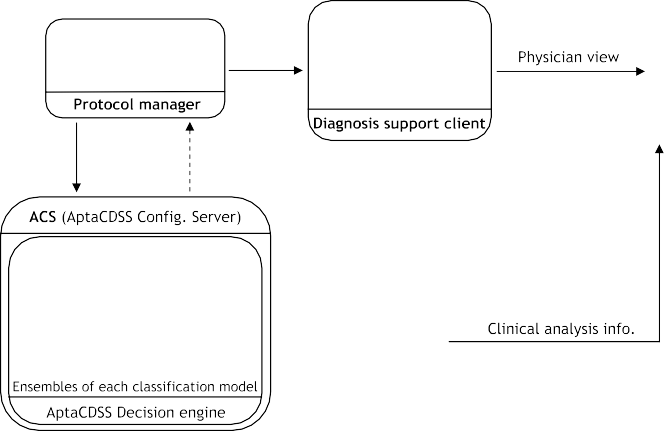
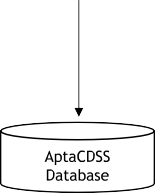
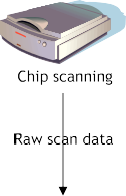
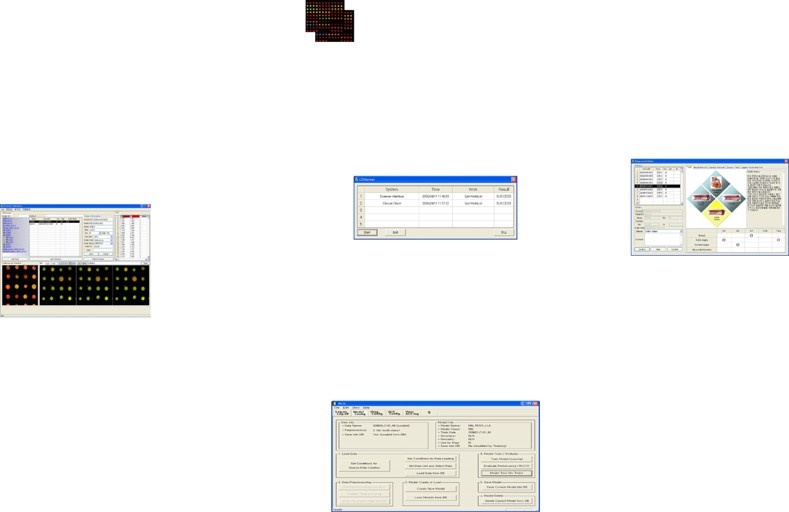


Fig. 1. The overall process ﬂow of the AptaCDSS-E. The system has four major components: ‘‘Scanner interface’’, ‘‘Protocol manager’’, ‘‘AptaCDSS conﬁguration sever (ACS)’’, and ‘‘Diagnosis support client’’. The ACS includes four classiﬁer ensembles of four diﬀerent classiﬁcation models for accurate clinical decision making. The solid lines indicate the ﬂow of data or system events and the dotted lines specify the ﬂow of diagnosis results or feedback information. The clinician’s analysis results can be delivered to the physician either directly or indirectly through the diagnosis support client.

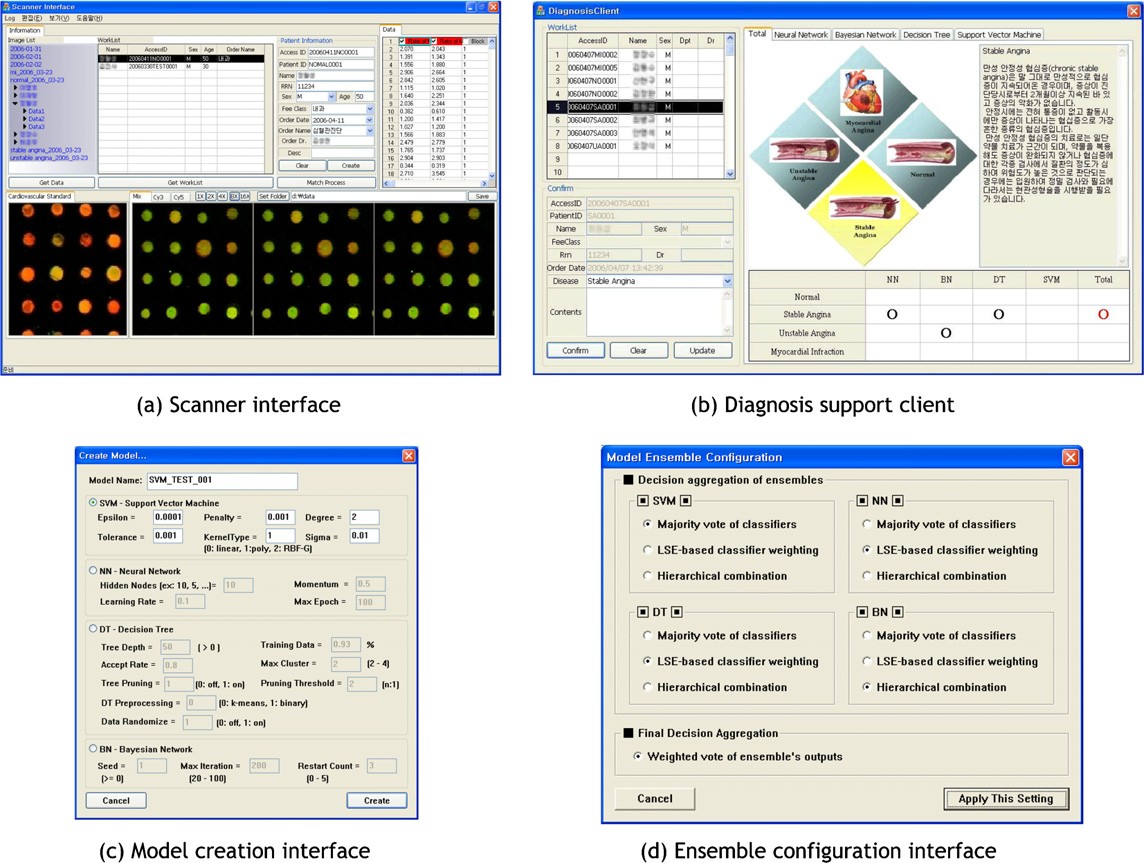


Fig. 2. The screenshots of AptaCDSS-E components. The scanner interface (a) reads raw scanner data ﬁles generated by chip scanner and converts it to a proper sample format. In the diagnosis support client (b), the Total tab combines the decision result of each classiﬁer and provides a simpliﬁed description of the current disease level of the patient. Each classiﬁer tab provides more detailed classiﬁer-speciﬁc or causal information to support the classiﬁer’s decisions. In the classiﬁer model creation interface (c), basic model parameters can be set. In the ensemble conﬁguration interface (d), the ensemble aggregation approach for each model can be conﬁgured. For ﬁnal decision aggregation, weighted voting of the ensemble’s output in terms of their prediction accuracies is provided. In this ﬁgure, several ﬁelds of conﬁdential patient information of scanner interface and diagnosis support client has been blurred for privacy.

overall process of diagnosis with the system and [Fig. 2](#_bookmark2) shows the interface examples of several components of the system.

* 1. *Scanner interface*

The scanner interface (SI) reads the original raw scanner generated data, composes patient chip sample data for each patient, and saves patient samples into the system database creating a new work list. In the SI, one can select speciﬁc ﬁelds of raw data to construct the patent sample. Users can also check and compare the status of the current chip expression image with standard sample images of cardio- vascular patients. For the development of AptaCDSS-E, the ‘‘ratio of median’’ ﬁeld of the original scanner data was selected to reduce negative eﬀects of outlier data points. [Fig. 2](#_bookmark2)a shows a screenshot of the SI.

* 1. *Protocol manager*

The protocol manager (PM), running in background, controls and meditates overall communications among the components by performing event scheduling and mes-

sage delivery. The communication part of the PM was implemented as a component (i.e., ActiveX) and combined with other elements of AptaCDSS-E. Each system compo- nent communicates by sending appropriate events to the server part of the PM. The server component also provides several monitoring functions of the component’s activity for system management.

* 1. *AptaCDSS conﬁguration server*

The AptaCDSS Conﬁguration Server (ACS) is the key part of AptaCDSS-E. The ACS performs diagnosis deci- sion making with pretrained classiﬁer ensembles of SVM, ANN, DT, and BN models. It also generates visualization information for the diagnosis support client. The ACS pro- vides a preprocessing function of patient samples to nor- malize an unprocessed initial sample dataset. Through the ACS, one can create basic decision models by setting model-speciﬁc parameters along with the proper conﬁgura- tion of ensemble constitution ([Fig. 2](#_bookmark2)c and d), train classiﬁer models with particular chip samples, test classiﬁer perfor- mance with diﬀerent data, and conﬁgure various settings for diagnosis and system logging.

* 1. *Diagnosis support client*

The diagnosis support client (DSC) provides integrated information to both physician and clinician. The disease progress levels of patient are classiﬁed into a total of four classes: ‘‘normal (NM)’’, ‘‘stable angina (SA)’’, ‘‘unstable

angina (UA)’’, and 14 ‘‘myocardial infarction (MI)’’. By

using the DSC, clinicians can analyze and select a set of

information gain is calculated with respect to entropy of each attributes, which deﬁned as

*c*

X

Entropyð*S*Þ— *pi*log2*pi*; ð1Þ

*i*¼1

Gainð*S*; *A*Þ¼ Entropyð*S*Þ— X j*S*mj Entropyð*S* Þ; ð2Þ

j*S*j

m

m2Valuesð*A*Þ

possible biomarker candidates for further detailed experi- mental validation, and physicians can make use of clini- cally analyzed information as supplementary diagnosis information. In addition to the supplementary information provided by clinicians, physicians can aided by the predic- tion results based on the set of prior patient samples. After the ﬁnal diagnosis is made by the physician, the physician can create and reﬂect feedback information to the system about unusual or exceptional cases for future reference by summarizing their opinions.

* 1. *Base classiﬁers*
     1. *Decision tree*

Decision tree induction is one of the most popular classi- ﬁcation methods. It builds a decision tree and classiﬁes the given data and has been successfully applied to a broad range of tasks. A decision tree is a tree in which each non- leaf node denotes a test on an attribute of cases, each branch corresponds to an outcome of the test, and each leaf node denotes a class prediction (see [Fig. 3](#_bookmark3)). To improve human readability, learned trees can also be re-represented as sets of if–then rules.

Decision trees select the most discriminant features based on the information gain at each stage when growing the tree structure. Consequently, a set of ordered features that make the largest contributions to successful classiﬁca- tion are obtained when classiﬁer training is ﬁnished. The

where *pi* is the proportion of outcomes belonging to class *i*, Values(*A*) is the set of all possible values for attribute *A*, and *S*m is the subset of for which attribute *A* has value m (i.e., *S*m = {*s S A*(*s*) = m}). In AptaCDSS-E, expression levels of proteins are discretized into one of four classes be- fore entropy calculation by applying *k*-means clustering- based preprocessing (*k* = 4) to generate comprehensible decision trees.

In this project, AptaCDSS-E utilized the C4.5 ([Quinlan,](#_bookmark35) [1993](#_bookmark35)) approach from among well-known decision tree induction algorithms for classifying CVD levels of interest and the values of protein expression as the attribute sets.

2 j

* + 1. *Neural network*

An ANN is a mathematical model consisting of a num- ber of highly interconnected processing elements organized into layers, the geometry and functionality of which have been inspired by that of the human brain. An ANN is trained with the available data samples to explore the rela- tion between inputs and outputs, so that one can reach the proper and accurate outputs when new data are added ([Simpson, 1990](#_bookmark39)). Multilayer perceptrons, a class of super- vised neural networks, is one of the most popular neural network models due to its clear architecture and compara- bly simple learning algorithm, and it is frequently used in MDSS ([Bishop, 1995; Ripley, 1996; Yan et al., 2006](#_bookmark27)).

For AptaCDSS-E, an MLP with a sigmoid function for node activation and standard back-propagation (BP)

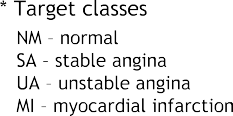
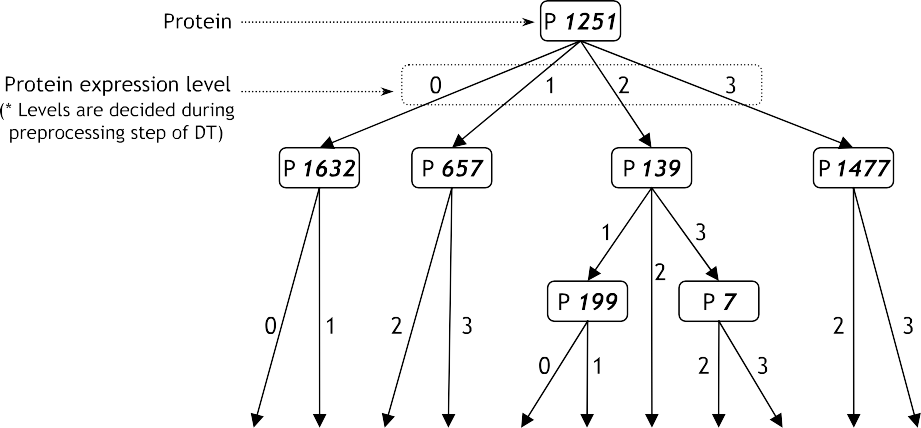


Fig. 3. A decision tree example for cardiovascular disease diagnosis. In this DT example, patient samples are classiﬁed into one of the four target classes by testing the expression value of seven marker proteins. The DT structure and checking markers for classiﬁcation of patient are obtained by training initial random structured DT with given training set. The marker proteins, such as P1251, are tested in which expression level they belong at each tree level (in this example, the proteins are classiﬁed into one of the four expression levels, 0, 1, 2, and 3) for patient diagnosis (classiﬁcation).

weight learning method were used. The BP algorithm is a widely used training procedure that adjusts the connection weights of the MLP ([Rumelhart, Hilton, & Williams,](#_bookmark35) [1986](#_bookmark35)). In BP, the error terms d*k* for each network output unit *k*, *ok*, and d*h* for each hidden unit *h*, *oh*, are calculated by

X

by the SVM, representing each category. For a linearly sep- arable binary classiﬁcation with an *n*-dimensional vector x*i* and the label of the class that vector *y* , i.e., fðx*i*; *y* Þg and

*yi* = {+1,—1}, the SVM separates the two classes of points

*i*

*i*

*i*¼1

*N*

using the classiﬁcation decision function *f*w,*b* = sign (w Æ x + *b*), where w is an input vector, x is an adaptive

d*k* ← *ok*ð1 — *ok*Þð*tk* — *ok*Þ and d*h* ← *oh*ð1 — *oh*Þ

*k*2outputs

*wkh*d*k*;

weight vector, and *b* is a bias. SVM ﬁnds the parameters w and *b* for the optimal hyperplane to maximize the geo- metric margin,

ð3Þ 2

wTw

where *tk* is the target value of unit *k*, *wkh* is the weight of

connection between the *k*th output unit and *h*th hidden

; subject to min

kwk

2 ; *yi*ðw · x*i* þ *b*Þ P þ1: ð5Þ

unit. The network weights are updated by

*wij* ¼ *wij* þ D*wij*; ð4Þ

where D*wij* = gd*ixij*. The output layer of MLP comprises of four nodes and each node corresponds to one cardiovascu- lar disease level of interest for prediction. The number of

nodes in the input layer varies according to the size of input

For the linearly non-separable case, the minimization problem needs to be modiﬁed to allow for the misclassiﬁ- cation of data points. A soft margin classiﬁer that allows

but penalizes errors by introducing slack variables n*l* as the measurement of violation of the constraints is ¼

*i* 1

repre-

sented by

feature vector determined by feature selection and the number of nodes in hidden layer is determined by user in- put (for AptaCDSS-E, we used 16 hidden nodes for three-

min

wTw

þ *C*

2

X

*N*

*i*¼1

*k*

n*i* ; *yi*ðw/ðx*i*Þþ *b*Þ P 1 — n*i*; ð6Þ

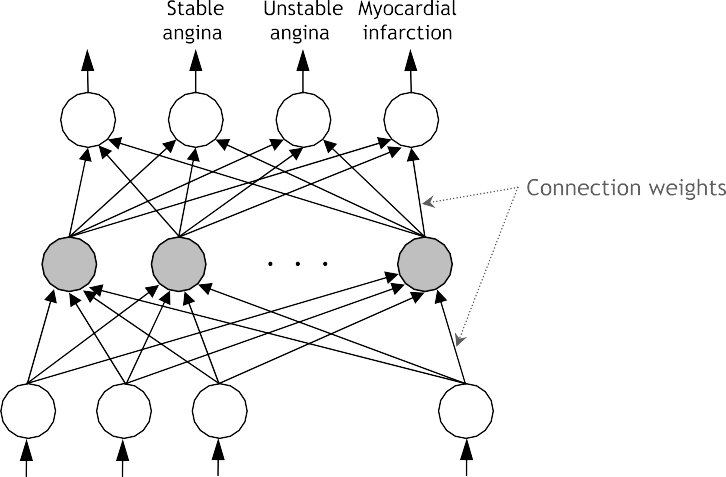
!

layered MLP). The architecture of the overall neural net- work classiﬁer is illustrated in [Fig. 4](#_bookmark5).

* + 1. *Support vector machine*

Support vector machines are an eﬀective binary data classiﬁcation method ([Vapnik, 1995](#_bookmark50)). The key idea of SVMs is the use of a mapping function which projects the given input feature space into a high dimensional fea- ture space to ﬁnd an optimal hyperplane having the largest margin of separation between diﬀerent classes with mini- mum error rate as shown in [Fig. 5](#_bookmark6).

SVMs use a portion of the data to train the system and ﬁnd several support vectors that represent the training data. These support vectors will be formed into a model



where *C* and *k* are used to weight the penalizing variables n*i*,

/(x*i*) is a non-linear function which maps the input space into a higher dimensional space (i.e., into a Hilbert space). This mapping can be represented as x*i* Æ x*j* /(x*i*) Æ / (x*j*) = *K*(x*i*, x*j*), where *K*(Æ) is a kernel function. Minimizing the ﬁrst term of Eq. [(6)](#_bookmark4) corresponds to minimizing the VC- dimension of the learning machine and minimizing the sec- ond term in Eq. [(6)](#_bookmark4) controls the empirical risk. The solution of this minimization problem can be found through a Wolfe dual problem with the Lagrangian method.

!

The SVM has several kernel functions that users can apply to solve diﬀerent problems. A proper inner product kernel function *K*(*xi* Æ *xj*) can solve certain linear insepara- ble problems without increasing the complexity of the



Fig. 4. The architecture of the three-layered MLP network as a base disease classiﬁer of AptaCDSS-E. For a given patient sample to diagnose, a vector of expression values of selected proteins is fed into the input layer. Each node of output layer corresponds to one target class of diagnosis and the class of the output node with maximum value is selected as a ﬁnal decision.

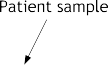




Fig. 5. The hyperplane-based linear separation of binary class data of SVM by feature space mapping. The SVM maximizes its margin of hyperplane in high dimensional feature space by ﬁnding the optimal hyperplane using support vectors. For one SVM, a total of six sub-SVMs are used in the manner of ‘‘1 to all’’ to perform four-class CVD patient classiﬁcation.

calculation and diﬀerent kernel functions are suited to diﬀerent problem types. The kernel function can be any function that satisﬁes Mercer’s theorem ([Mercer, 1909](#_bookmark56)); however, the most popularly used kernel functions are the linear, polynomial, and radial basis functions, and sigmoid kernels. For AptaCDSS-E, we have chosen the polynomial kernel.

* + 1. *Bayesian network*

A Bayesian network ([Cooper & Herskovits, 1992; Heck-](#_bookmark31) [erman, Geiger, & Chickering, 1995](#_bookmark31)) is a graphical model

that represents dependency relationships among variables

of interest. It is represented as an annotated directed acy-

discrete random variables where each X*i* may take on val- ues from a ﬁnite domain is the pair *B* = *G*, *L* . The *G* is a DAG whose nodes correspond to the random variables X1,... , X*n*, and whose edges represent direct dependencies between the variables. The graph structure *G* encodes the following set of independence statements: each variable X*i* is independent of its non-descendents, given its parent in *G*. Standard arguments ([Pearl, 1988](#_bookmark62)) shows that any dis- tribution *P* that satisﬁes the independence statements encoded in the graph *G* can be factored as

*P* ð*X* ; ... ; *X* Þ¼ Y *P* ð*X* jPa Þ; ð7Þ

h i

1

*n*

*i*

*i*

*n*

*i*¼1

clic graph (DAG) encoding probabilistic relationships among distinctions of concern in an uncertain-reasoning problem. The nodes or the vertices of the DAG represent the random variables in the network while the edges con- necting the vertices represent the causal inﬂuence of one node on the other. Each node of graph has a probability table representing probabilistic relations with other con- nected nodes. By using the given network structure, prob- ability table, and some observations of partial variables, an inference for other unobserved variables can be made. Formally, a BN for a given ﬁnite set U = {X1,... , X*n*} of

where Pa*i* denotes the parents of X*i* in *G*.

The second component of the BN, *L*, is a set of condi- tional probabilities between the variables in *G*. The prob- lem of training a BN can be stated as a task of ﬁnding an optimal network *Bs* that best matches the given training set *D* = {u1,... , u*N*} i.e., to ﬁnd a network that maximizes *P*(*Bs D*) = *P*(*Bs*, *D*) *P*(*D*).

j j

The learning processes of a BN include structure learn- ing of *G* and parameter learning of *L*. The structure learn- ing, the optimization problem in the space of the DAGs, ﬁnds an appropriate graph structure for the given data

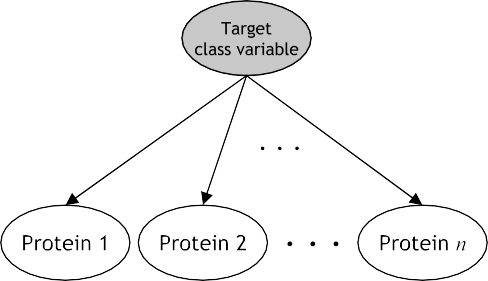
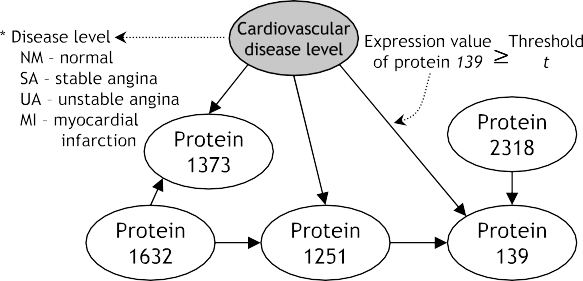
 



Fig. 6. The structure of a na¨ıve Bayes model as an initial BN and the example structure result of BN structure learning. The BN of the system starts network structure learning from this na¨ıve Bayes model structure which has *n* edges connecting target class variable and feature (protein) variables to learn variable dependency for a given data set (a). After structure and parameter learning, the learned BN model is used to diagnose given test samples by deciding one class of the target classes of target class variable (b). In the resulting BN, each edge represents causality between nodes by representing ‘‘underexpression’’ (the expression value of a node < threshold *t*) or ‘‘overexpression’’ (the expression value of a node P threshold *t*).

from all possible graph constitutions. Since network struc- ture ﬁnding is known to be an NP-hard problem ([Hecker-](#_bookmark36) [man et al., 1995](#_bookmark36)), various heuristics have been applied in structure learning such as greedy search, greedy search with restart, best-ﬁrst search, and simulated annealing, etc. In our study, ‘‘greedy search with random restart’’ method which is a simple but robust heuristic approach was used to resolve the problem of local optimum convergence. We used the na¨ıve Bayes classiﬁer structure of [Fig. 6](#_bookmark7)a as an initial network structure of network learning with this search strategy. This approach modiﬁes its initial simple BN structure by adding, deleting, and switching the direc- tions of the edge in consecutive order and selects a network with highest score among networks obtained by several repeated trials. The ﬁtness of a network structure was eval- uated by ‘‘Bayesian Dirichlet and likelihood equivalence’’ (BDe) score metric ([Heckerman et al., 1995](#_bookmark36)).

After the network structure learning, conditional proba- bilities of each variable of the obtained network for given parent nodes are calculated in the extended framework of BDe by calculating suﬃcient statistics from given data with ﬁxed priors.

1. Ensemble of classiﬁers
   1. *Need for a classiﬁer ensemble*

The complexity and subtlety of microarray expression patterns between CVD patients and normal samples may increase the chance of misclassiﬁcation when a single clas- siﬁer is used because a single classiﬁer tends to cover pat- terns originating from only part of the sample space. Therefore, it would be beneﬁcial if multiple classiﬁers could be trained in such a way that each of the classiﬁers covers a diﬀerent part of the sample space and their classiﬁcation results were integrated to produce the ﬁnal classiﬁcation. Moreover, this combination can reduce the variance of estimation errors and improve the overall classiﬁcation accuracy ([Shin & Markey, 2006](#_bookmark40)).

Ensemble algorithms such as bagging, boosting, or ran- dom forests improve the classiﬁcation performance by associating multiple base classiﬁers to work as a ‘‘commit- tee’’ for decision-making and any supervised learning algo- rithm can be used as a base classiﬁer of ensemble ([Bauer &](#_bookmark21) [Kohavi, 1999](#_bookmark21)). Ensemble algorithms not only increase the classiﬁcation accuracy, but also reduce the chances of over- training since the committee avoids a biased decision by integrating the diﬀerent predictions from the individual base classiﬁers. The concept of combining classiﬁers into ensembles ﬁrst appeared in work by [Nilson (1965)](#_bookmark60) (further described in [Sharkey, 1999](#_bookmark41)), and then extensive studies started in the 1990s.

For this reason, AptaCDSS-E adopted the ensemble approach to generate enhanced results by grouping a set of classiﬁers of each SVM, ANN, DT, and BN. In this sec- tion, we will describe the classiﬁer combination approaches adopted by AptaCDSS-E.

* 1. *Why ensemble works better*

An ensemble of classiﬁers is a set of classiﬁers whose individual decisions are combined in some way (typically weighted or unweighted voting) to classify new examples. It is known that ensembles are often much more accurate than the individual classiﬁers that make them up. An ensemble can be more accurate than its component classi- ﬁers only if individual classiﬁers disagree with one another ([Hansen & Salamon, 1990](#_bookmark36)).

For example, for an ensemble of three classiﬁers:

{*h*1, *h*2, *h*3} and we consider a new case x. If the three clas- siﬁers are identical, then when *h*1(x) is wrong, *h*2(x) and *h*3(x) are also wrong. However, if the errors made by the classiﬁers are uncorrelated, then when *h*1(x) is wrong, *h*2(x) and *h*3(x) might be correct, so that a majority vote correctly classiﬁes x. More precisely, if the error rates of *L* hypotheses *h*‘ are all equal to *p* < 1/2 and if the errors are independent, then the probability that the majority vote is wrong is the area under the binomial distribution where more than *L*/2 hypotheses are wrong. Of course, if the indi- vidual hypotheses make uncorrelated errors at rates exceeding 0.5, then the error rate of the voted ensemble increases as a result of the voting. Hence, the key to suc- cessful ensemble methods is to construct individual classiﬁ- ers with error rates below 0.5 whose errors are at least somewhat uncorrelated.

* 1. *Construction of classiﬁer ensemble*

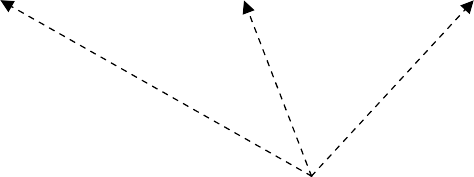
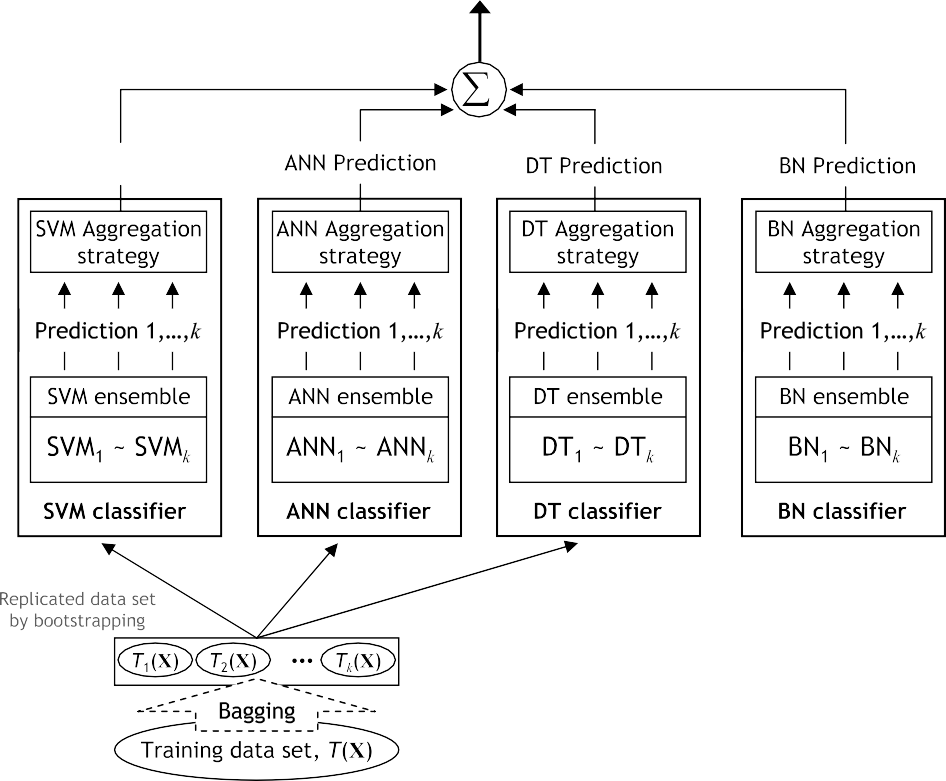
Many approaches for constructing an ensemble of clas- siﬁers have been proposed. The most important thing in constructing a classiﬁer ensemble is to make each individ- ual classiﬁer diﬀerent from the other classiﬁers as possible. This requirement can be met by using diﬀerent training sets for diﬀerent classiﬁers. In AptaDSS-E, one of the represen- tative methods, bagging ([Breiman, 1996](#_bookmark28)), is used to satisfy this requirement of classiﬁer diversity.

In a bagging, classiﬁers are trained independently via a bootstrap method and then they are aggregated by an appropriate combination strategy. Bootstrapping gener- ates *K* replicas {*Tk*(X) *k* = 1,..., *K*} of training data by repeated random re-sampling with replacement from the given training data *T*(X) = {(*xi*; *yy*) *i* = 1,..., *N*}. As a result, each example in the given training set may appear repeatedly or not at all in any particular replica training set. Then, each replicated training set is used to train a certain classiﬁer of an ensemble.

j

j

To achieve maximal diversity of ensembles, we can con- struct ensembles with diﬀerent classiﬁcation models. But in this case, it is not easy to compare diﬀerent classiﬁer mod- els because the diﬀerence comes from model-speciﬁc char- acteristics of the models in the ensemble. Furthermore, we need a well-deﬁned objective measure to compare fairly a set of diﬀerent kind of models. Hence, we construct an ensemble for each classiﬁcation method with *k* homoge- neous classiﬁers, but make them diﬀerent as much as





Fig. 7. The overall architecture of the decision making part of AptaCDSS-E with multiple classiﬁer ensembles (here we used a total of four ensembles of four diﬀerent classiﬁcation models). Each ensemble is constituted with several classiﬁer models of each classiﬁcation method. The training data are augmented by bagging and fed to each classiﬁer member of ensembles. Final decision is decided by weighted majority vote of each ensemble’s decision with respect to their training accuracies.

possible by setting initial factors randomly such as weights, structures, and probabilities.

* 1. *Classiﬁer aggregation*

After training the classiﬁers of each model group, we need to aggregate independently trained classiﬁers of each group into an appropriate combination method. We con- sidered two types of model combination approaches such as linear (the majority voting and LSE-based weighting) and non-linear (the double-layer hierarchical grouping) combination method ([Kim, Pang, Je, Kim, & Bang, 2003](#_bookmark42)).

* + 1. *Majority voting*

One simplest method of classiﬁer combination is major- ity voting. For *fk* (*k* = 1,..., *K*), a decision function of the *k*th classiﬁer in the classiﬁer ensemble, and *cj* (*j* = 1,..., *C*), a label of *j*th class, the ﬁnal decision of an ensemble *f*vote(x) for a given test data x with majority voting is decided by

*f*voteðxÞ¼ arg max t*j*; ð8Þ

*j*

where t*j* is the number of classiﬁers whose decisions are known to *j*th class and deﬁned by t*j* = *c*(*k*, *j*), where *c*(*k*, *j*) is 1 if *fk*(x) = *cj* and 0, otherwise.

P

* + 1. *Least squared error (LSE)-based classiﬁer weighting*

The LSE-based weighting of classiﬁers treats several classiﬁers in the classiﬁer ensemble with diﬀerent weights. The weights of diﬀerent classiﬁers are assigned in propor- tional to their classiﬁcation accuracies. For *fk*(*k* = 1,..., *K*), a decision function of the *k*th classiﬁer in the classiﬁer ensemble which trained with a replica of training data *T k* X *x*0*i*; *yi*0 *i* 1; ... ; *N* , the weight vector w can be

ð Þ¼ fð Þj ¼ g

obtained by w*E* = A—1y, where A = (*fi*(x*j*))*K*·*N*, and y =

(*yj*)1·*N*. Then, the ﬁnal decision of the classiﬁer ensemble

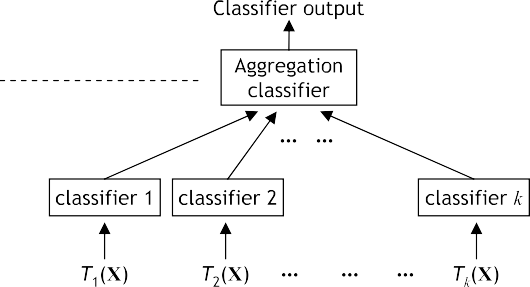


Fig. 8. Hierarchical combination of classiﬁers. The classiﬁer’s decision outputs in the lower layer are fed into aggregation classiﬁer in the upper layer and ﬁnal decision of the ensemble is made by this classiﬁer.

Fig. 9. The whole experimental steps of the aptamer chip-based disease level classiﬁcation process.



for a given test data vector x with LSE-based weighting is decided by

*f*LSEðxÞ¼ signðw · ½ð*fi*ðxÞÞ*k*×1]Þ: ð9Þ

This weight-based linear combination is also used to combine the decision results of each classiﬁer ensemble with respect to their accuracy on the training data to make the ﬁnal decision as shown in [Fig. 7](#_bookmark10).

* + 1. *Hierarchical combination*

In hierarchical combination, an additional classiﬁer is used to aggregate the outputs of classiﬁers of the ensemble. So, this combination consists of a double-layer of classiﬁers where the outputs of several classiﬁers in the lower layer feed into an aggregation classiﬁer in the upper layer ([Fig. 8](#_bookmark11)).

For *fk*(*k* = 1,..., *K*), a decision function of the *k*th clas- siﬁer in the classiﬁer ensemble, and a decision function of the aggregating classiﬁer *F*, the ﬁnal decision function of the classiﬁer ensemble *f*HC(x) for given test data x with the double-layer hierarchical combination is given by

*f*HCðxÞ¼ *F* ð*f*1ðxÞ; *f*2ðxÞ; ... ; *fk*ðxÞÞ; ð10Þ

where *k* is the number of classiﬁers in the ensemble.

* 1. *Making the ﬁnal decision*

The ﬁnal decision in [Fig. 7](#_bookmark10) is decided by combining out- puts of all ensembles taking accuracy-based weighted majority vote (i.e., use their training accuracies as their weights). Then, the ﬁnal class *c*ﬁnal among the possible tar- get classes (*C*, *C* = 0: Normal, 1: SA, 2: UA, 3: MI) is decided by

in the vote by giving some advantage to ensembles with re- spect to the preference of each classiﬁcation method.

1. Experimental results and discussion

The experimental steps of aptamer chip-based disease level classiﬁcation with multiple classiﬁers are summarized in [Fig. 9](#_bookmark13). The steps in category A were performed by the data supplier, and in this research we performed the steps with solid border in category B and C with the Apta- CDSS-E. The ﬁnal experimental veriﬁcation of discovered possible biomarkers will be conducted in future work.

* 1. *Data sets*

The AptaCDSS-E performs clinical decision support task of cardiovascular disease by analyzing aptamer chip data, which were produced from the patient’s blood sam- ples. The advantages of using blood samples include: blood is readily accessible and less expensive to obtain than many other procedures. The disease analysis of AptaCDSS-E is performed on blood-derived products, particularly on serum which is the ﬂuid that remains after clothing proteins are removed from plasma.

Besides the CVD data, we used three additional disease data sets, which include pulmonary complaints, tuberculo- sis disease, and general cancer collections to overcome the data insuﬃciency problem and evaluate the generalized classiﬁcation accuracy of the system for other diseases. [Table 1](#_bookmark15) shows the statistics of CVD and other disease sam- ples used in this study.

Table 1

The statistics of four data sets

*n* Disease Feature

Sample

Number of target

*c*final ¼ arg max X *Ii*;*c*ð*wi* þ c*i*Þ; ð11Þ

*c*2*C*

*i*¼1

where *n* is the number of classiﬁer ensembles, *wi* is the

dimension size class

Cardiovascular disease

3000 66 4 (normal, SA,

(CVD) UA, MI)

Pulmonary complaints 3000 95 2 (normal,

weight of *i*th ensemble, and *Ii*,*c* is the indicator of *i*th ensem-

ble, which has 1 if the output class of ensemble is equal to *c*

and 0, otherwise. The c*i* in Eq. [(11)](#_bookmark14) is an advantage vari-

(PC)

Tuberculosis disease (TBD)

complaints)

1000 27 2 (normal, tuberculosis)

able, predetermined variable by the user, preventing a draw

General cancer (GC) 1000 54 2 (normal, cancer)

* 1. *Preprocessing*

We constructed chip data using the ‘‘ratio of median’’ ﬁeld of the original scanner generated ﬁle to minimize the negative properties of outlier data points. Also, the set of

control spots for each chip sample were removed and the missing values are ﬁlled with the median value of the sam- ple. Next, the data were transformed by applying logarithm base 2 and a ratio-based normalization method is applied to adjust the means of the samples to zero. By applying this

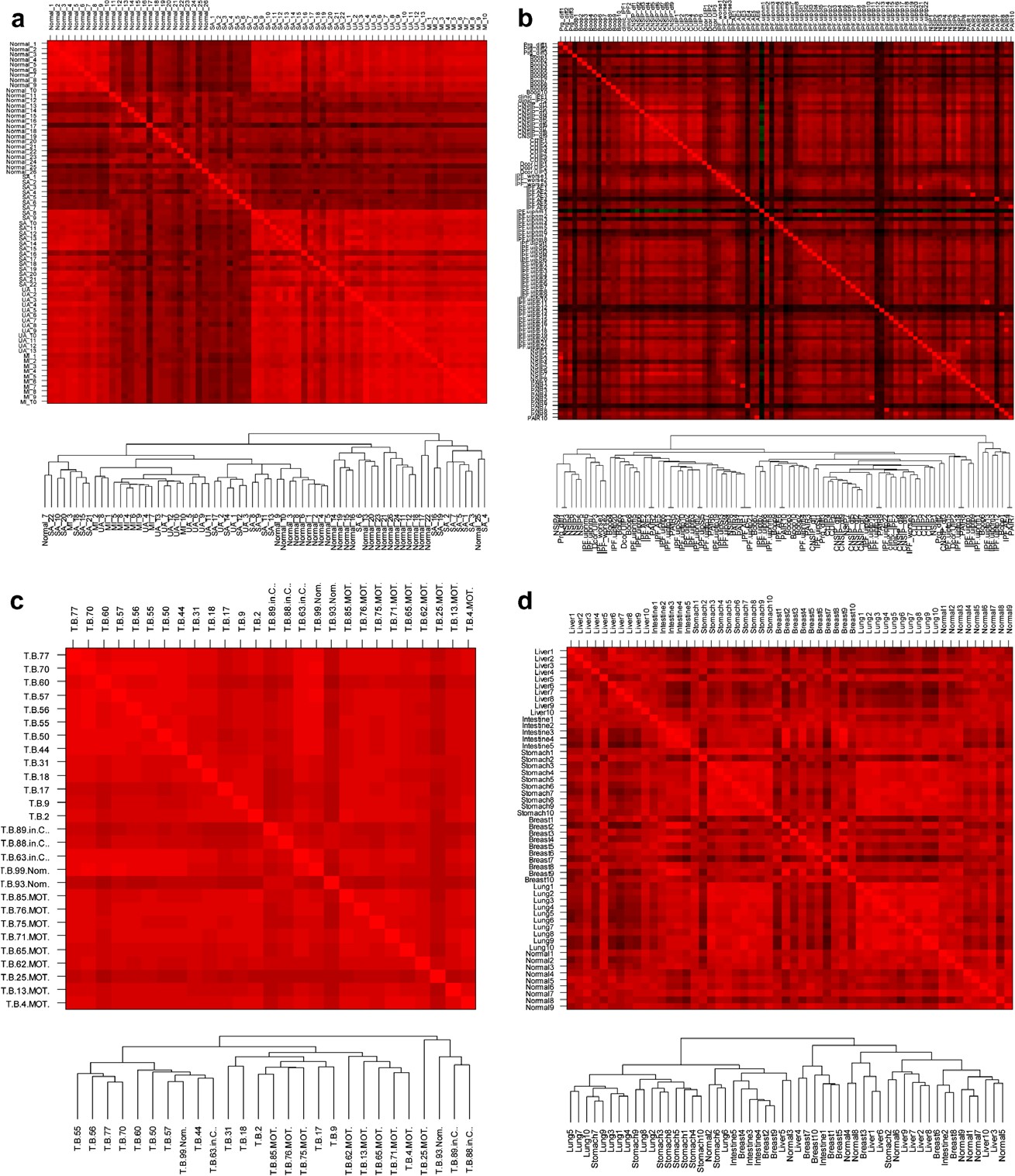


Fig. 10. The quality of the four data sets with respect to their correlation and hierarchical clustering results without feature selection. In the correlation matrix, the red dots indicate ‘‘positive correlation’’, the green dots indicate ‘‘negative correlation’’, and the black ones indicate ‘‘no-correlation’’ between samples. The samples of each data set are clustered moderately by hierarchical clustering with average linkage. The ‘‘general cancer’’ data set includes cancer samples of liver, lung, intestine, breast, stomach, and nine normal CVD samples for binary class classiﬁcation (for ‘‘normal vs. cancer’’ comparison). a: Cardiovascular disease, b: Pulmonary disease, c: Tuberculosis, d: General cancer.

ratio-based normalization we removed the laser channel diﬀerence which may occur in chip scanning process.

* 1. *Feature selection*

For the feature selection, the dimension of each disease data set is reduced by applying analysis of variance (ANOVA) and we selected the top 250 proteins according to their signiﬁcance score (*p*-value) to build a ﬁnal classiﬁer inputs. [Fig. 10](#_bookmark16) shows the quality of four disease data sets with respect to their correlation analysis (Pearson correla- tion) and hierarchical sample clustering results with their full features. The quality of these data sets is reﬁned by dis- carding non-informative features according to their signif- icance (*p*-value). By applying this feature selection, we could also reduce the complexities of data processing in each classiﬁcation model.

* 1. *Results*

[Table 2](#_bookmark17) shows the classiﬁcation accuracy of each classi- ﬁer and of each classiﬁer ensemble for diﬀerent ensemble constitution methods for each data set. The classiﬁcation (prediction) performance was measured by *k*-fold cross-val- idation with *k* = 10 to alleviate the insuﬃciency of samples. For ensemble-based model, each ensemble is trained with the data set augmented by bagging described in Section [3.3](#_bookmark9). In the case of the single classiﬁer-based prediction, the SVM performed best for all data sets and the ANN ranked second among the four diﬀerent classiﬁers. The prediction intervals of the classiﬁers were about 6.8% at least and 12.2% at most. Presumably, the prediction accuracy for ‘‘Tuberculosis disease’’ data was relatively low due to the

small sample size and poor quality of data.

For the ensemble classiﬁer-based prediction, the SVM and ANN performed very well for all data sets similar to

the single classiﬁer case. Especially, the ANN achieved the best prediction accuracy for ‘‘Pulmonary complaints’’ data for all ensemble aggregations and SVM achieved the best prediction accuracy for ‘‘General cancer’’ data. Inter- estingly, the BN ensemble with LSE-based weighting aggregation method was the best classiﬁer for ‘‘Tuberculo- sis diseases’’. The maximum prediction interval of overall ensemble method was about 7% ([86.39, 93.42], for the ensemble with LSE-based weighing for ‘‘GC’’ data) and the minimum was about 2.1% ([89.53, 91.64], for the ensemble with majority vote for ‘‘TBD’’ data). The hierar- chical classiﬁer combination showed the best performance among the three aggregation methods showing prediction accuracy intervals between about 5.5% ([90.36, 95.87],

for ‘‘CVD’’ data) and 2.4% ([90.97, 93.38], for ‘‘TBD’’

data).

By utilizing DT and BN, we obtained decision support information, including causalities among sample features, which can be represented in a human readable and easy to understand structure such as rules or causality networks. [Fig. 11](#_bookmark18) shows a simple BN example with 10 nodes for car- diovascular disease diagnosis trained with CVD data of [Table 1](#_bookmark15).

In [Fig. 11](#_bookmark18), the ﬁnal decision probabilities of four classes in the class node (four classes; 0, 1, 2, and 3 for NM, SA, UA, and MI, respectively) are decided by setting the pro- tein node’s expression value to a binary value according to whether the measures expression is greater or less than the sample’s median. For the given sample data (i.e., each protein’s binary expression level), the example BN diagno- ses the current sample as ‘‘Normal (NM)’’ class (see the probability bar chart of class node in [Fig. 11](#_bookmark18)).

In the BN of [Fig. 11](#_bookmark18), the ﬁnal diagnosis decision is made by choosing the class of maximum probability value in the class node. The probabilities of each target class in the class node are calculated by multiplying the highest conditional

Table 2

The classiﬁcation accuracy of single and ensemble-based classiﬁers with diﬀerent classiﬁer aggregation method for four diﬀerent data sets Classiﬁer composition Classiﬁer aggregation Classiﬁer Accuracies for each data set

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | | | CVD | PC | TBD | GC |
| Single classiﬁer | – | SVM | 84.31 ± 1.2 | 82.92 ± 1.5 | 76.32 ± 1.3 | 81.64 ± 1.6 |
|  |  | ANN | 80.82 ± 1.1 | 81.64 ± 0.9 | 73.94 ± 1.2 | 80.21 ± 1.4 |
|  |  | DT | 72.69 ± 1.6 | 70.68 ± 1.3 | 69.54 ± 1.7 | 70.01 ± 1.1 |
|  |  | BN | 78.95 ± 2.1 | 77.51 ± 1.2 | 71.43 ± 2.6 | 70.39 ± 1.5 |
| Ensemble-based classiﬁer | Majority voting | SVM | 92.82 ± 1.0 | 94.31 ± 0.0 | 91.64 ± 1.3 | 93.11 ± 1.1 |
|  |  | ANN | 93.49 ± 0.9 | 94.55 ± 0.9 | 90.21 ± 0.5 | 91.01 ± 0.7 |
|  |  | DT | 91.03 ± 1.0 | 90.66 ± 0.7 | 89.53 ± 1.4 | 87.41 ± 1.1 |
|  |  | BN | 92.01 ± 0.7 | 92.34 ± 1.1 | 90.08 ± 0.9 | 89.96 ± 0.8 |
|  | LSE-based weighting | SVM | 93.08 ± 0.7 | 94.66 ± 0.8 | 90.62 ± 0.9 | 93.42 ± 0.8 |
|  |  | ANN | 94.12 ± 0.6 | 94.98 ± 0.7 | 89.08 ± 0.7 | 90.71 ± 1.1 |
|  |  | DT | 90.03 ± 0.5 | 89.83 ± 0.9 | 90.57 ± 0.9 | 86.39 ± 0.8 |
|  |  | BN | 90.17 ± 0.4 | 90.09 ± 0.7 | 92.37 ± 0.5 | 88.95 ± 1.2 |
|  | Hierarchical combination | SVM | 95.87 ± 0.3 | 95.67 ± 0.2 | 92.68 ± 0.6 | 94.31 ± 0.3 |
|  |  | ANN | 94.32 ± 0.5 | 95.72 ± 0.2 | 93.38 ± 0.4 | 93.18 ± 0.7 |
|  |  | DT | 90.36 ± 0.9 | 92.19 ± 0.5 | 91.04 ± 0.8 | 88.83 ± 0.7 |
|  |  | BN | 92.51 ± 0.4 | 93.11 ± 0.2 | 90.97 ± 0.5 | 91.53 ± 0.4 |

Accuracies in bold indicate maximum values of each conﬁguration.

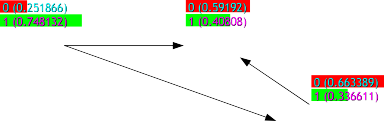
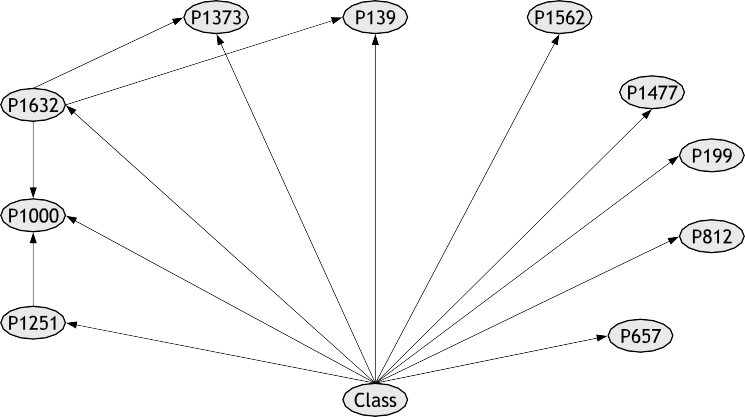


Fig. 11. The 10-node BN for CVD diagnosis generated by the BN model of AptaCDSS-E. Each node represents one protein selected from 3000 input protein features (250 proteins after preprocessing), and class node represents the ﬁnal decision of this BN. The probability bar chart of each node represents its cumulative probability quantiﬁed by BN training with training samples (DT results are not shown in this paper).

probabilities of all nodes for a given sample of data. [Fig. 12](#_bookmark19) shows the conditional probability tables for each node of BN in [Fig. 11](#_bookmark18). The probability values of each conditional probability table are calculated by BN learning with the CVD data in [Table 1](#_bookmark15).

* 1. *Discussion*

The results of the experiment ([Table 2](#_bookmark17)) show that an improvement in prediction accuracy of more than

about 10% has been achieved by applying the ensemble method. In particular, hierarchical combination of classiﬁ- ers showed great accuracy improvements, implying that it is one of the desirable classiﬁer combination methods in ensemble construction. Generally, SVMs, the current state-of-the-art classiﬁer, and ANNs, the most widely adopted model for clinical diagnosis application, achieved relatively higher accuracies than other classiﬁers.

Although SVMs and ANNs achieved the best perfor- mance for most data sets, it is not easy to understand how

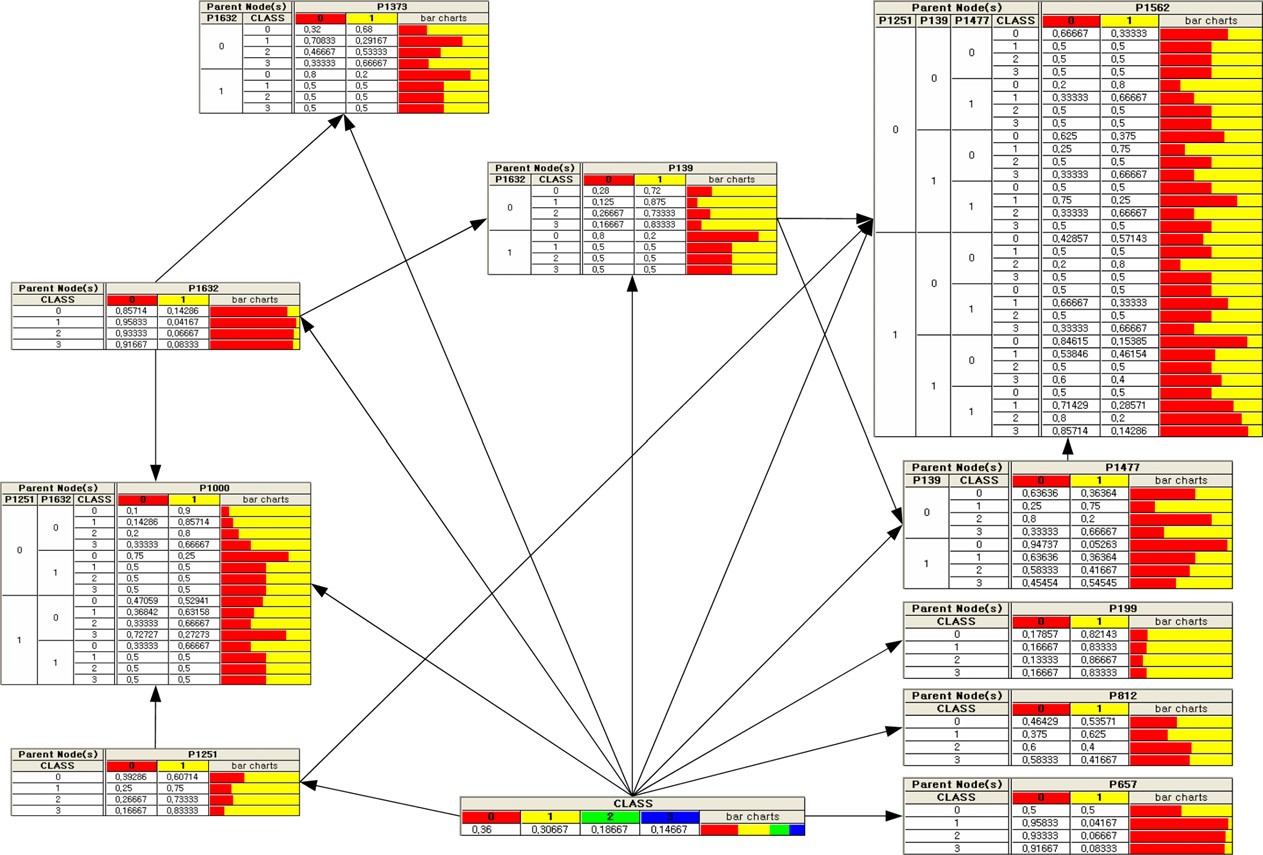


Fig. 12. The conditional probability tables of the BN in [Fig. 11](#_bookmark18). For each node, the size of table increases as the number of parent nodes increase.

they produced diagnosis results because they use non-linear feature mappings and weight compositions. For this reason, they are often referred to as ‘‘black box’’ models. Conse- quently, these models are not appropriate for generating diagnosis support information that can be used by physi- cians and clinicians to help in their decision making.

In AptaCDSS-E, we adopted DTs and BNs to resolve this diﬃculty and generate decision support supplementary information displayed by DSC. Generally, DTs generate human readable decision rules and BNs generate causality networks, which can be easily understood by humans. The BN of [Fig. 11](#_bookmark18) shows the causalities of the major 10 proteins selected from the total of 3000 proteins (250 proteins after preprocessing) for the diagnosis of cardiovascular disease. These selected proteins can be regarded as a set of possible biomarkers for CVD diagnosis and can be conﬁrmed as ‘‘real’’ biomarkers after further experimental veriﬁcation. Moreover, this information can be used by clinicians to design new clinical trials by utilizing those proposed possi- ble biomarkers for disease diagnosis.

To summarize, the advantage of using the proposed sys- tem is such that physicians can have practical aids in their daily diagnosis with relatively high accuracy and clinicians can ﬁnd meaningful ‘‘real’’ biomarkers by investigating the results produced by AptaCDSS-E.

However, even though we adopted an ensemble-based classiﬁer approach and bagging as a data augmentation strategy to boost prediction accuracies, data sampling tech- niques cannot overcome coverage limitations inherent to the data set from which the samples are drawn. If the data set does not represent the underlying probability distribu- tion of the population of interest, then even the most sophisticated feature selection based on sampling tech- niques will end up with an extremely biased subset of fea- tures. In case of TBD data, the size of samples was too small, and it seems that the sample data sets did not con- tain the sample characteristics appropriate for classifying their classes. Consequently, the poor quality of samples led to degradation of overall prediction accuracy for this data set for all classiﬁers. Therefore, securing more micro- array chip samples with relatively good quality and reﬂect- ing the underlying characteristics of a disease of concern is one of the most important issues in achieving improved and generalized classiﬁcation accuracy.

1. Conclusions

We have presented a classiﬁer ensemble-based clinical decision support system called AptaCDSS-E for disease level prediction with aptamer biochip data. The system employs four diﬀerent machine learning classiﬁers, com- bines the prediction results of each classiﬁer in an ensemble machine, and generates supplementary information for dis- ease diagnosis. The system was trained with four diﬀerent disease data sets consisting of 242 cases including cardiovas- cular disease and the data sets were augmented by bagging for classiﬁer ensemble training. The experimental result

with cross-validation shows that the proposed system pre- dicts the level of diseases with relatively high accuracy (>94%) and small prediction diﬀerence intervals (<6%), showing its usefulness in support of clinical decision making for diagnosis. In particular, causality information among the major 10 proteins for cardiovascular disease diagnosis was found by the system as a candidate set of possible bio- markers, which now require further clinical veriﬁcation.

Acknowledgements

This work was supported by the Korea Science and Engineering Foundation (KOSEF) through the National Research Lab. Program funded by the Ministry of Science and Technology (No. M10400000349-06J0000-34910) and supported by the Korea Research Foundation Grant funded by the Korean Government (MOEHRD) (KRF- 2006-511-D00355). The authors would like to thank Byoung-Hee Kim, Je-Keun Rhee, Min-Oh Heo, Young- Jin Park, and Min-Hyeok Kim for the construction of AptaCDSS platform, and developers and researchers of GenoProt Co. Ltd. for data preparation.

References

Balla, J. I., Iansek, R., & Elstein, A. (1985). Bayesian diagnosis in presence of preexisting disease. *Lancet, 325*(8424), 326–329.

Bauer, E., & Kohavi, R. (1999). An empirical comparison of voting classiﬁcation 37 algorithms: Bagging, boosting, and variants. *Machine Learning, 36*(1–2), 105–139.

Baxt, W. G. (1990). Use of an artiﬁcial neural network for data analysis in clinical decision making: The diagnosis of acute coronary occlusion. *Neural Computing, 2*(4), 480–489.

Baxt, W. G. (1995). Application of artiﬁcial neural networks to clinical medicine. *Lancet, 346*(8983), 1135–1138.

Bishop, C. M. (1995). *Neural networks for pattern recognition*. Oxford, UK: Oxford University Press.

Brasil, L. M., de Azevedo, F. M., & Barreto, J. M. (2001). Hybrid expert system for decision supporting in the medical area: Complexity and cognitive computing. *International Journal of Medical Informatics, 63*(1–2), 19–30.

Breiman, L. (1996). Bagging predictors. *Machine Learning, 24*(2), 123–140.

CDC’s Report 1. Accessed: 29.05.06. [http://www.cdc.gov/nccdphp/](http://www.cdc.gov/nccdphp/overview.htm) [overview.htm](http://www.cdc.gov/nccdphp/overview.htm).

CDC’s (2006). Accessed: 29.05.06. [http://www.cdc.gov/nccdphp/publica-](http://www.cdc.gov/nccdphp/publications/aag/cvh.htm) [tions/aag/cvh.htm](http://www.cdc.gov/nccdphp/publications/aag/cvh.htm).

Chae, Y. M., Park, K. E., Park, K. S., & Bae, M. Y. (1998). Development of medical decision support system for Leukemia management. *Expert Systems with Applications, 15*, 309–315.

Conforti, D., & Guido, R. (2005). Kernel-based support vector machine classiﬁers for early detection of myocardial infarction. *Optimization Methods and Software, 20*(2–3), 401–413.

Cooper, G. F., & Herskovits, E. (1992). A Bayesian method for the induction of probabilistic networks from data. *Machine Learning, 9*(4), 309–347.

Cortes, C., & Vapnik, V. (1995). *Support-vector networks machine learning*

(pp. 237–297). Boston, MA: Kluwer Academic Publisher.

Dybowski, R., & Gant, V. (1995). Artiﬁcial neural networks in pathology and medical laboratories. *Lancet, 346*(8984), 1203–1207.

Ellenius, J., & Groth, T. (2000). Transferability of neural network-based decision support algorithms for early assessment of chest-pain patients. *International Journal of Medical Informatics, 60*(1), 1–20.

Geurts, P., Fillet, M., de Seny, D., Meuwis, M. A., Malaise, M., Merville,

M. P., et al. (2005). Proteomic mass spectra classiﬁcation using decision tree based ensemble methods. *Bioinformatics, 21*(14), 3138–3145.

Gu¨ ler, N. F., & Koc¸er, S. (2005). Use of support vector machines and neural network in diagnosis of neuromuscular disorders. *Journal of Medical Systems, 29*(3), 271–284.

Gu¨ ven, A., & Kara, S. (2006). Classiﬁcation of electro-oculogram signals using artiﬁcial neural network. *Expert Systems with Applications,* *31*(1), 199–205.

Hansen, L. K., & Salamon, P. (1990). Neural network ensembles. *IEEE* *Transactions on Pattern Analysis and Machine Intelligence, 12*, 993–1001.

Harris, N. L. (1990). Probabilistic belief networks for genetic counseling.

*Computer Methods and Programs in Biomedicine, 32*(1), 37–44.

He, J., Hu, H. J., Harrison, R., Tai, P. C., & Pan, Y. (2006).

Transmembrane segments prediction and understanding using support vector machine and decision tree. *Expert Systems with Applications,* *30*(1), 64–72.

Heckerman, D., Geiger, D., & Chickering, D. (1995). Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning, 20*(3), 197–243.

Kaplan, B. (2001). Evaluating informatics applications-clinical decision support systems literature review. *International Journal of Medical Informatics, 64*(1), 15–37.

Kim, H.-C., Pang, S., Je, H.-M., Kim, D., & Bang, S. Y. (2003).

Constructing support vector machine ensemble. *Pattern Recognition, 36*(12), 2757–2767.

Kim, M.-J., Min, S.-H., & Han, I. (2006). An evolutionary approach to the combination of multiple classiﬁers to predict a stock price index. *Expert Systems with Applications, 31*(2), 241–247.

Li, Y. C., Liu, L., Chiu, W. T., & Jian, W. S. (2000). Neural network modeling for surgical decisions on traumatic brain injury patients. *International Journal of Medical Informatics*, 57(1), 1–9.

Liu, T.-F., Sung, W.-K., & Mittal, A. (2006). Model gene network by semi-ﬁxed Bayesian network. *Expert Systems with Applications, 30*(1), 42–49.

MacDowell, M., Somoza, E., Rothe, K., Fry, R., Brady, K., & Bocklet,

A. (2001). Understanding birthing mode decision making using artiﬁcial neural networks. *Medical Decision Making, 21*(6), 433–443. Majumder, S. K., Ghosh, N., & Gupta, P. K. (2005). Support vector machine for optical diagnosis of cancer. *Journal of Biomedical Optics,*

*10*(2), 24–34.

Mangalampalli, A., Mangalampalli, S. M., Chakravarthy, R., & Jain, A.

K. (2006). A neural network based clinical decision-support system for eﬃcient diagnosis and fuzzy-based prescription of gynecological diseases using homoeopathic medicinal system. *Expert Systems with Applications, 30*(1), 109–116.

Mendyk, A., & Jachowicz, R. (2005). Neural network as a decision support system in the development of pharmaceutical formulation – focus on solid dispersions. *Expert Systems with Applications, 28*(2), 285–294.

Mercer, T. (1909). Functions of positive and negative type and their connection with the theory of integral equations. *Transaction of London Philosophy Society (A), 209*, 415–446.

Murphy, C. K. (2001). Identifying diagnostic errors with induced decision trees. *Medical Decision Making, 21*(5), 368–375.

Nilson, N. J. (1965). *Learning machines: Foundations of trainable pattern classiﬁers*. New York: McGraw-Hill.

Ohlsson, M. (2004). WeAidU – A decision support system for myocardial perfusion images using artiﬁcial neural networks. *Artiﬁcial Intelligence in Medicine, 30*(1), 49–60.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Mateo, CA: Morgan Kaufmann Publishers.

Prados, J., Kalousis, A., Sanchez, J. C., Allard, L., Carrette, O., & Hilario,

M. (2004). Mining mass spectra for diagnosis and biomarker discovery of cerebral accidents. *Proteomics, 4*(8), 2320–2332.

Qu, Y., Adam, B.-L., Yasui, Y., Ward, M. D., Cazares, L. H., Schellhammer, P. F., et al. (2002). Boosted decision tree analysis of surface-enhanced laser desorption/ionization mass spectral serum proﬁles discriminates prostate cancer from noncancer patients. *Clinical Chemistry, 48*(10), 1835–1843.

Quinlan, J. R. (1993). *C4.5: Programs for machine learning*. Los Altos, CA: Morgan Kaufmann Publishers.

Ripley, B. D. (1996). *Pattern recognition and neural networks*. Cambridge: Cambridge University Press.

Rumelhart, D. E., Hilton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature, 323*, 533–536.

Sadeghi, S., Barzi, A., Sadeghi, N., & King, B. (2006). A Bayesian model for triage decision support. *International Journal of Medical Informat- ics, 75*(5), 403–411.

Schubert, F., Mu¨ ller, J., Fritz, B., Lichter, P., & Eils, R. (2003). Understanding the classiﬁcation of tumors with a support vector machine: A case-based explanation scheme. *Proceedings of the German conference on bioinformatics (GCB 2003)*, Neuherberg/Garching, 12– 14 October (pp. 123–127).

Sharkey, A. (1999). Multi-net systems. In A. J. C. Sharkey (Ed.), *Combining artiﬁcial neural nets – Ensemble and modular multi-net systems* (pp. 1–30). Berlin: Springer-Verlag.

Shin, H., & Markey, M. K. (2006). A machine learning perspective on the development of clinical decision support systems utilizing mass spectra of blood samples. *Journal of Biomedical Informatics, 39*(2), 227– 248.

Simpson, P. K. (1990). *Artiﬁcial neural systems: Foundations paradigms applications and implementations*. Elmsford, NY: Pergamon Press.

Stockwell, D. R. B. (1993). LBS: Bayesian learning system for rapid expert system development. *Expert Systems with Applications, 6*(2), 137– 147.

Tung, K.-Y., Huang, I.-C., Chen, S.-L., & Shih, C.-T. (2005). Mining the generation xers’ job attitudes by artiﬁcial neural network and decision tree – Empirical evidence in Taiwan. *Expert Systems with Applications, 29*(4), 783–794.

Tu¨ rkoglu, I., Arslan, A., & Ilkay, E. (2002). An expert system for diagnosis of the heart valve diseases. *Expert Systems with Applications, 23*(3), 229–236.

Vapnik, V. (1995). *The nature of statistical learning theory*. New York: Springer-Verlag.

Veropoulos, K., Cristianini, N., & Campbell, C. (1999). The application of support vector machines to medical decision support: A Case Study. In *Proceedings of the ECCAI advanced course on artiﬁcial intelligence (ACAI 1999)*, 5–16 July 1999, Chania, Greece.

West, D., Mangiameli, P., Rampal, R., & West, V. (2005). Ensemble strategies for a medical diagnostic decision support system: A breast cancer diagnosis application. *European Journal of Operational Research, 162*(2), 532–551.

West, D., & West, V. (2000). Model selection for a medical diagnostic decision support system: A breast cancer detection case. *Artiﬁcial Intelligence in Medicine, 20*(3), 183–204.

Won, Y., Song, H., Kang, T. W., Kim, J., Han, B., & Lee, S. (2003).

Pattern analysis of serum proteome distinguishes renal cell carcinoma from other urologic diseases and healthy persons. *Proteomics, 3*(12), 2310–2316.

Yan, H.-M., Jiang, Y.-T., Zheng, J., Peng, C.-L., & Li, Q.-H. (2006). A

multilayer perceptron-based medical decision support system for heart disease diagnosis. *Expert Systems with Applications, 30*(2), 272–281.

Yang, S., & Browne, A. (2004). Neural network ensembles: Combining multiple models for enhanced performance using a multistage approach. *Expert Systems, 21*(5), 279–288.