Hierarchical Probabilistic Support Vector Machine for Detecting Cardiovascular Diseases

Mubo Chen, Binbin Fu, Taichun Tang, Jiali Ma, and Mingchui Dong

Abstract—32 Hemodynamic parameters (HDPs) derived from sphygmogram (SPG) and 5 physiological parameters (PPs) are widely used for cardiovascular diseases (CVDs) detection. These parameters are divided to groups naturally for diagnostic usage, which conforms to doctors’ clinical diagnosis procedure. The number and type of HDP&PP in groups are varied according to detecting different CVDs, in another word the grouping is disease-oriented, which leads to a bottleneck problem: how to construct a hierarchical high-efficient classifier to diagnose CVDs based on grouped HDPs & PPs? To tackle such a formidable problem, a hierarchical classifier HPSVM based on support vector machine with probabilistic outputs (PSVM) is proposed. Such formed classifier has good generalization ability with unique global solution even with small training dataset. It also conforms to the doctors’ hierarchical diagnosis procedure and reduces the deduction complexity with high diagnostic accuracy. Site-measured datasets obtained from Beijing Changping Chinese Medicine Hospital are used for testing and the results verify the prospect of this technology with higher than 90% accuracy in detecting three typical and frequently encountered CVDs.

Index Terms—Classification and regression tree, disease-oriented grouping, hierarchical classifier, quadratic programming, support vector machine with probabilistic outputs.

I. INTRODUCTION

Cardiovascular diseases (CVDs) have emerged as the top health killer in both urban and rural areas in most of the countries. Thereamong, coronary heart disease (CHD), hypertension (HT) and hyperlipemia (HL) are three typical and frequently encountered CVDs harming elderly patients’ health with high morbidity and mortality [1]. For this reason, CVDs detection has drawn a great deal of attention [2]-[4] for decades.

CVDs can be detected by diagnosing vital signs such as electrocardiography (ECG), Echocardiography and/or Sphygmogram (SPG). Comparing with ECG and Echocardiography, SPG is competitive and widely used in e-home healthcare usage for its non-invasive and easy acquisition. Based on SPG signals, hemodynamic parameters (HDPs) are derived by using the model of elastic cavity, which is capable of revealing cardiovascular health status and variation tendency [5], [6]. Hence, hard efforts of exploring HDPs have been made by researchers for CVDs detection through applying artificial intelligence technology.

Concretely, decision tree can fast deduce conclusion from HDPs according to exclusively expressed medical knowledge [7], but calculations might get extremely complex while handling uncertain values; support vector machine shows high accuracy in detecting CVDs [8], but has disadvantage of time consuming and contradicting to the doctors’ clinical diagnosis procedure since it is one-layer classifier. In practice, doctornormally ranks all parameters and selects specified ones with most pertinence as diagnostic basis in accordance with different diseases, which makes the reasoning procedure representing “hierarchically” character. There have been some literatures relating to the research of hierarchical diagnosis system, such as hierarchical fuzzy neural networks (HFNNs) [6] performed with promising learning capability and robustness, but has the well-known disadvantage of lacking exclusive explanation to the diagnostic result. Thus, [9] proposed using a genetic algorithm (GA) to extract accurate and comprehensive fuzzy IF-THEN rules and gained good performance. However, it is proved by [10]. The HDP&PP categorization method should be disease-oriented, whereas generic categorization method is applied in existing common hierarchical CVDs diagnosis classifiers [6], [9]. In other words, a hierarchical and self-adapting classifier concerning applying disease-oriented HDP&PP groups hierarchically to diagnose different CVDs is necessary and needed. Virtually, there is no relevant literature about intelligent classification method solving this bottleneck problem so far.

Thus, in this research, a hierarchical probabilistic support vector machine (HPSVM) for CVDs detection using HDP&PP groups hierarchically is proposed which reduces complexity of inference computation, increases diagnostic accuracy and speeds as well as conforms to doctor’s clinical diagnosis procedure. Besides, a specific HDP&PP categorization method CARTCM [10] was applied for realizing disease-oriented HDP&PP grouping. By taking the diversity of HDP&PP categorization into account, the CVDs diagnostic accuracy of hierarchical classifier is highly increased.

II. CATEGORIZATION OF HDPs AND PPS

Before categorization of HDPs & PPs, data representation in mathematical form is needed and inevitable. As shown in

Manuscript received March 8, 2014; revised May 8, 2014. This work was supported in part by the Research Committee of University of Macau under Grant MYRG184 (Y2-L3)-FST11-DMC, and Science and Technology Development Fund (FDCT) of Macau S.A.R under Grant 018/2009/A1, respectively.

Mubo Chen, Binbin Fu, Jiali ma, and Mingchui Dong are with the Department of Electrical and Computer Engineering, faculty of Science and Technology, University of Macau, Macau S.A.R., China (e-mail: chenmubo@ieee.org, binbinfu@umac.mo, ariespleo51@gmail.com, jimohanqiu@hotmail.com, mcdong@umac.mo)

Taichun Tang is with the Department of Software Engineering, Faculty of Science and Technology, University of Macau, Macau S.A.R., China (e-mail: tangtao09@163.com).
Table I, the on-site measured medical records consist of ID number, patients’ personal information, patients’ PPs and HDPs as well as doctors’ diagnostic results. Each HDP&PP record is denoted as $x_i = [x_{i1}, x_{i2}, ..., x_{iM}]^T$ ($i = 1, 2, ..., M$), where $M$ is the total number of patients’ records and $N$ is the total number of HDP&PP. Specifically, $x_j$ is the $j$th record’s $i$th HDP&PP, where $j = 1, 2, ..., N$. Thus all patients’ HDP&PP records are denoted as matrix $X_{N \times M}$:

$$X_{N \times M} = [x_1, x_2, ..., x_M]$$

(1)

The diagnostic result $Y_{D \times M}$ obtained by medical inference is denoted as:

$$Y_{D \times M} = Y_{D1} Y_{D2} ... Y_{DM}$$

(2)

where $D$ is the total number of CVD types and specifically $y_{di} \in \{-1, +1\}$ is diagnostic result of the $i$th record’s $d$th CVD. $y_{di} = -1$ represents the patient measured with the $i$th record does not have the $d$th CVD, while $y_{di} = +1$ means the patient measured with the $i$th record has the $d$th CVD. In this research, totally 37 parameters including 32 HDPs and 5 PPs (the acronyms of HDPs and PPs can be found in [6]) are employed to diagnose three types of typical and frequently encountered CVDs (CHD, HT, HL), thus $N = 37$ and $D = 3$.

<table>
<thead>
<tr>
<th>ID</th>
<th>NAME</th>
<th>A</th>
<th>H</th>
<th>MST</th>
<th>CHD</th>
<th>HT</th>
<th>HL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Patient 1</td>
<td>58</td>
<td>155</td>
<td>15.07</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>2</td>
<td>Patient 2</td>
<td>55</td>
<td>172</td>
<td>14.40</td>
<td>+1</td>
<td>+1</td>
<td>+1</td>
</tr>
<tr>
<td></td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

For CVD detection, the aim is to infer the diagnostic result $Y_{D \times M}$ from $X_{N \times M}$, which is expressed as:

$$Y_{D \times M} = S_{D \times N} \cdot X_{N \times M}$$

(3)

where $S_{D \times N}$ is a nonlinear mapping which can be realized by SVM, HFNNs or so on. Obviously, using all 37 HDPs & PPs at once to detect CVDs is paradoxical, influencing diagnostic accuracy, high computation-consuming and unforformable to doctors’ normal clinical diagnosis procedure. Therefore, in this research, our proposed CARTCM is applied before CVDs detection.

The CARTCM shown in Fig. 1 includes two parts:

1) $D$ classification and regression trees using $X$ plus $Y$ are constructed. Corresponding to the $d$th CVD, the importance index of each HDP&PP denoted as $V_{dji}(d = 1, 2, ..., D$ and $j = 1, 2, ..., N$) is computed. Specifically, in terms of the $d$th CVD, $V_{dji}$ is the importance index of the $i$th HDP&PP, which is calculated by the popular and useful measurement Gini Index [10]. Then all HDPs & PPs importance indexes are normalized in $(0, 1]$. In detail, $V_{dji}^{\text{normalized}}$ is the normalized importance index of the $i$th HDP&PP in terms of the $d$th CVD. After that, corresponding to different types of CVDs, all the HDPs & PPs are ranked from the most to the least important according to their normalized importance indexes;

2) HDPs and PPs are dynamically categorized into three groups by a thresholding method which is defined in (4) according to pre-detected type of CVD. Here, the normalized importance indexes of HDPs and PPs are gradually declined from Group $d_1$ to Group $d_3$ (most, medium, least important group), where $d = 1, 2, ..., D$.

$$G(V_{dji}^{\text{normalized}}) = \begin{cases} \text{Group } d_1, & V_{dji}^{\text{normalized}} \in (\alpha_1, 100]\% \\ \text{Group } d_2, & V_{dji}^{\text{normalized}} \in (\alpha_2, \alpha_1]\% \\ \text{Group } d_3, & V_{dji}^{\text{normalized}} \in [0, \alpha_2]\% \end{cases}$$

(4)

$\alpha_1$, $\alpha_2$ are tunable parameters which are set empirically as 60% and 20% respectively in this research. As an example, for detecting the $d$th CVD, a HDP or PP with normalized importance index $V_{dji}^{\text{normalized}} = 90\%$ should be categorized into Group $d_1$ due to $90\% \in (\alpha_1, 100\%)$. For more details about CARTCM, please refer to [10].

III. HIERARCHICAL PROBABILISTIC SVM METHOD

For detecting the $d$th CVD, 37 HDPs & PPs are divided into three groups (Group $d_1$, Group $d_2$, Group $d_3$) from the most important to the least important according to CARTCM as mentioned above, and then a disease-oriented hierarchical classifier HPSVM with three PSVM layers is constructed as shown in Fig. 2 (here, PSVM is a sub-classifier based on SVM which could output probabilistic estimate). Firstly, HDPs & PPs in Group $d_1$ are selected and fed to Level $d_1$PSVM and it outputs the diagnostic result with probability estimate. If the probability is larger than the pre-defined threshold $\theta_{d_1}$, then the inference will stop and output the final diagnostic result, which is the best case. Otherwise, HDPs & PPs in next group combined with previous group (Group $d_2$) will be chosen to feed into the next Level $d_2$PSVM. The same inference mechanism is repeated for Group $d_3$. In the worst
case, all 37 HDPs & PPs (parameters in Group\(_{1,23}\), Group\(_{2,23}\), Group\(_{3,23}\) are combined) are used and fed to Level\(_{1,23}\) PSVM, thus outputing the final diagnostic result. Obviously, PSVM with probabilistic estimation is the key module to be constructed for HPSVM. Next depicts its details.

PSVM developed by Platt [11] is a promising method of machine learning based on the theory of Vapnik-Chervonenkis dimension and the principle of structural risk minimum. Comparing with traditional SVM [12], PSVM has an advantage of probabilistic estimation outputs by using the Sigmoid-fitting method. For this reason, PSVM is adopted in HPSVM. First, an optimal hyper-plane is searched, which separates the set of training vectors belonging to two separated classes as well as makes the margin of the separation beside the optimal hyper-plane maximum while ensuring the accuracy of correct classification. In this research, such a hyper-plane is used for detecting CVDs. In details, for detecting the \(d^1\) CVD for a new-come patient with HDP&PP record \(x_i\), the separating hyper-plane with the target functions is expressed as \(f(x) = (w^*\cdot x) + b\), where \(\cdot\) is the inner product operation; \(w\) and \(b\) represent the normal vector and the bias respectively. If \(f(x_i) > 0\), thus set \(y_{di} = +1\), it means this patient has the \(d^1\) CVD disease, otherwise set \(y_{di} = -1\). \(w\) and \(b\) can be determined by solving following quadratic programming (QP):

\[
\min \frac{1}{2} (w^T \cdot w) + C \sum_{i=1}^{M} \xi_i
\]

\[
s.t. \quad y_{di} \left( (w^\star \cdot x_i) + b \right) \geq 1 - \xi_i, \quad \xi_i \geq 0; \quad d = 1, 2, ..., \; \text{D}; \; i = 1, 2, ..., \; \text{M}.\]

where the meanings of \(M\) and \(D\) are same as what explained in previous section, \(C\) is the penalty parameter and \(\xi_i\) \((i = 1, 2, ..., \; M)\) is the slack variable. The optimization mission is a QP task thus can be well solved by adopting the Sequential Minimal Optimization algorithm [13]. The posterior probabilistic estimation is approximated by a sigmoid function [11]:

\[
\Pr(y_{di} = 1|x_i) = P_{y_{di} = 1}(f) = \frac{1}{1 + e^{-f}} \quad (6)
\]

For a new-come patient with HDP&PP record \(x_i\), the separating hyper-plane with the target functions is expressed as \(f(x) = (w^*\cdot x) + b\), where \(\cdot\) is the inner product operation; \(w\) and \(b\) represent the normal vector and the bias respectively. If \(f(x_i) > 0\), thus set \(y_{di} = +1\), it means this patient has the \(d^1\) CVD disease, otherwise set \(y_{di} = -1\). \(w\) and \(b\) can be determined by solving following quadratic programming (QP):

\[
\min \frac{1}{2} (w^T \cdot w) + C \sum_{i=1}^{M} \xi_i
\]

\[
s.t. \quad y_{di} \left( (w^\star \cdot x_i) + b \right) \geq 1 - \xi_i, \quad \xi_i \geq 0; \quad d = 1, 2, ..., \; D; \; i = 1, 2, ..., \; M.\]

where the meanings of \(M\) and \(D\) are same as what explained in previous section, \(C\) is the penalty parameter and \(\xi_i\) \((i = 1, 2, ..., \; M)\) is the slack variable. The optimization mission is a QP task thus can be well solved by adopting the Sequential Minimal Optimization algorithm [13]. The posterior probabilistic estimation is approximated by a sigmoid function [11]:

\[
\Pr(y_{di} = 1|x_i) = P_{y_{di} = 1}(f) = \frac{1}{1 + e^{-f}} \quad (6)
\]

The performances of HPSVM classifier compared with HFNNs are shown in Table III. The prevalent C-support vector classification model and Radial Basis Function kernel are used in this research. Besides, 8-fold validation is applied, which means 357 records are used as training dataset while 51 records as testing dataset for every time (totally 8 times). It is observed that HPSVM shows higher Acc, SE and PPV except the performance of PPV for detecting HL is relatively lower. The lower PPV indicates that some healthy people might be wrongly deduced as HL patients occasionally. However, as this method is proposed for e-home healthcare usage, all abnormal cases are friendly suggested to be further confirmed by competent hospital, thus such PPV is acceptable. Table III also shows that HPSVM method gains higher Acc, SE and PPV than HFNNs proposed in [6]. The reasons can be concluded as (1) HPSVM applies the disease-oriented CARTCM categorization method to divide HDPs & PPs into groups for detecting different kinds of CVDs more specifically, while HFNNs use a generic categorization method; (2) the solution of HPSVM is global and unique while HFNNs may suffer from multiple local minima due to the disadvantage of Back Propagation Neural Networks (BPNNs); (3) HPSVM uses structural risk minimization with good generalization ability while HFNNs use empirical risk minimization and may suffer from over-fitting; (4) HPSVM has good a performance even with relative small training dataset.
It is also observed that HPSVM shows higher time efficiency. The comparison of training time between HPSVM and HFNNs indicates that HPSVM has higher degree of computing efficiency. In our experiment, for CHD, Level1, Level2, PSVM or Level1, Level2, Level3 PSVM are needed, which is the same case with HL, while for detecting HT, only Level1 PSVM is required. In essence, when detecting one kind of CVD, firstly the most important HDP&PP group and Level1 (d = 1, 2, …, D) PSVM are used. If Level1 PSVM outputs the diagnostic result with high probability which is larger than threshold θd, then the diagnosis is ended. Otherwise, the HDP&PP Groupd combining with Group1 and Level1 PSVM will be used and the above inference procedure is repeated until obtaining the final diagnostic result (empirical thresholds θ1 and θd corresponding to different kind of CVDs are shown in Table IV). Obviously HPSVM conforms to doctor’s clinical diagnosis procedure and shows higher time efficiency in comparison of applying all HDPs & PPs as input in one-layer classifier.

### TABLE II: DISEASE-ORIENTED HDP & PPS CATEGORIZATION

<table>
<thead>
<tr>
<th>Disease</th>
<th>Group1</th>
<th>Group2</th>
<th>Group3</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHD (d=1)</td>
<td>A, SWI, W, TBV, SL, CCP, TBVR</td>
<td>BV, CBMB, SPR, DP, SV, MHR, PR, TPR, CO, PP, CI, BEK, H, MIST, MDP, VER, MRT, EWK, CMBR</td>
<td>VPK, FEK, SP, HOV, PAR, MSP, MAP, AC, HOI, PAP, PAWP</td>
</tr>
<tr>
<td>HT (d=2)</td>
<td>VPK, PP, SP, A, FEK, SV</td>
<td>EWK, SI, MISP, MST, VER, MDP, BEK, HOI, MAP, CMBR, TBVR, MRT</td>
<td>CCP, CI, PAR, PAWP, CO, PAP, BV, W, PR, CBMB, HOV, DP, TBV, SWI, MHR, AC, H, SPR, TPR</td>
</tr>
<tr>
<td>HL (d=3)</td>
<td>DP, MHR, CCP, TBVR, A, SPR</td>
<td>BEK, SV, TPR, PR, CO, SI, BV, PAR, CBMBV, MDP, AC, MRT, MRT, HOV, SWI, VPK, W</td>
<td>HOI, EWK, CMBR, MSP, H, VER, CI, MAP, PAP, PP, PAWP, SP, FEK</td>
</tr>
</tbody>
</table>

### TABLE III: PERFORMANCE BETWEEN HFNNS AND HPSVM

<table>
<thead>
<tr>
<th>Disease</th>
<th>HFNNs (Acc (%))</th>
<th>HFNNs (SE (%))</th>
<th>HFNNs (PPV (%))</th>
<th>HPSVM (Acc (%))</th>
<th>HPSVM (SE (%))</th>
<th>HPSVM (PPV (%))</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHD</td>
<td>46.15</td>
<td>96.00</td>
<td>100.00</td>
<td>94.47</td>
<td>93.38</td>
<td>91.99</td>
</tr>
<tr>
<td>HT</td>
<td>53.85</td>
<td>84.85</td>
<td>100.00</td>
<td>96.88</td>
<td>94.82</td>
<td>90.00</td>
</tr>
<tr>
<td>HL</td>
<td>53.92</td>
<td>85.71</td>
<td>100.00</td>
<td>91.35</td>
<td>98.21</td>
<td>99.81</td>
</tr>
</tbody>
</table>

### TABLE IV: SETTING OF THRESHOLDS θd AND θd2 CORRESPONDING TO VARIOUS CVDs

<table>
<thead>
<tr>
<th>Thresholds</th>
<th>CHD (d = 1)</th>
<th>HT (d = 2)</th>
<th>HL (d = 3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>θd</td>
<td>0.9</td>
<td>0.9</td>
<td>0.9</td>
</tr>
<tr>
<td>θd2</td>
<td>0.9</td>
<td>0.9</td>
<td>0.6</td>
</tr>
</tbody>
</table>

### V. CONCLUSION

A HPSVM is proposed to diagnose three typical and frequently encountered CVDs hierarchically, in which CARTCM is adopted to carry disease-oriented HDP&PP categorization. The number and type of HDPs and PPs in each of three groups are determined automatically according to importance and relevance of HDP&PP in detecting various CVDs. With a hierarchical inference mechanism, the HDP&PP groups from the most important to the least are selectively fed to HPSVM’s different layers depending on the probabilistic estimation outputs. The testing results and comparisons show out its higher diagnostic accuracy and prove the validity of presented methodology.

### REFERENCES


Mubo Chen is currently pursuing MSc degree in ECE Department of University of Macau, Macau. His current research is oriented to R&D a hybrid intelligent system to diagnose cardiovascular disease. He received Bachelor degree in CS in 2012 from Guangdong University of Foreign Studies. His research interests include machine learning, applying AI technology in various engineering and biomedical applications etc.

Binbin Fu is currently pursuing PhD degree in ECE Department of University of Macau, Macau. Her research interests focus on embedded system, biomedical signal processing and artificial intelligence, towards the goal of synthesizing them for wearable/porable monitoring with computerized risk stratification.
Taichun Tang is currently a postgraduate student in Software Engineering, University of Macau, Macau, China. He received Bachelor degree in Computer Science and Technology in 2012 from Shihezi University, Sinkiang, China. His current research is oriented to R&D 3-Rank CVD intelloigent diagnosis. His main research interests are machine learning, application of AI in biomedical engineering and similarity and complexity analysis.

Jia Li Ma received the B.Sc. degree in electrical and information engineering from Tianjin University, Tianjin, China in 2011. Currently, she is pursuing master degree in electrical and computer engineering at University of Macau. Her research interests include biomedical signals compression and construction of embedded-link e-health systems.

Minghui Dong is a full professor and PhD supervisor of ECE Department of University of Macau, Automation Department of Tsinghua University, China and Professor of YanTai University, China. He received MSc degree in EEE in 1975 at Tsinghua University, China, Visiting Scholar in EEE in 1981 at Rome University, Italy.

His main research interests are AI and its application in biomedical engineering, CIMS, fault diagnosis, AI in vision-to-text, voice-to-text machine translation etc.