Support Vector Machines to Define and Detect Agitation Transition

George E. Sakr, *Member*, *IEEE*, Imad H. Elhajj, *Senior Member*, *IEEE*, and Huda Abou-Saad Huijer

Abstract—The need to automate the detection of agitation and the detection of agitation transition for dementia patients is a significant facilitator for caregivers. This research aims at detecting the transitional phase toward agitation, as well as agitation detection of subjects, using soft computing techniques that do not require supervision beyond the training phase. Three vital signs are monitored: Heart Rate (HR), Galvanic Skin Response (GSR), and Skin Temperature (ST). These measures are fed into two proposed SVM architectures which are based on the definition of a new confidence measure: "Confidence-Based SVM" and "Confidence-Based Multilevel SVM." Results show very high detection accuracy of agitation and agitation transition, a quick adaptation to the subject, and a strong correlation between the physiological signals monitored and the emotional states of the subjects. Another challenge that is successfully addressed in this paper is the ability to train the classifier on a limited group of subjects, and then test it on subjects not belonging to the training group. The result is a learning algorithm that is "Subject-Independent."

Index Terms—Agitation detection, agitation transition detection, support vector machines, confidence.

1 Introduction

Dementia is an acquired syndrome of decline in at least one function of the cognitive system, such as language, attention, problem solving, visuospatial, and memory, and is sufficient to disturb the social and occupational life of an alert person [1]. Dementia is very common in the aging population, but can also hit at an early age and at any stage of adulthood. A very common sign of an advanced stage of dementia is disorientation in time or in place for people suffering from this syndrome. In coming years, the number of older patients potentially accessing the health care system is expected to double, increasing from 35 million in 2010 to 72 million by 2030 [2]. This population boom, a 78 percent increase, will result in 1 in every 5 Americans being over the age of 65 by the year 2030 [2]. The rapidly growing elderly population brings new health care issues into sharp view, particularly that of caring for elders with Alzheimer's Disease, which is the leading cause of dementia [2]. It is in the long term care setting that the effects of dementia are particularly felt, as the disorder results in negative behavioral symptoms in 54 percent of patients [2]. This is primarily due to agitation and can interrupt patient care and frustrate caregivers.

Caring for elders with dementia is psychologically demanding and can result in psychiatric symptoms of caregiver "burnout" [3]. This also has serious implications. The loss of skilled health care providers will undoubtedly be felt in the often short-staffed world of long-term care. In order to control costs, provide optimal patient care, and prevent the burnout of skilled professionals, efficient methods of agitation measurement and management must be implemented to aid nursing staff in their efforts.

The rest of the paper is organized as follows: Section 2 discusses previous techniques of agitation and emotion detection. Section 3 introduces SVM, VC dimension, the proposed confidence measure, and the resulting detection algorithms. Then the experimental results are presented in Section 4.

2 Related Work

Emotion detection has been shown to be an essential tool in developing machine intelligence [4]. In general, an emotion detection experiment can be treated through measuring different types of biological signals or by using image processing techniques for facial expression analysis [5]. This paper uses biological signals for emotion detection, which, in general, has four main challenges:

- The choice of the biological signals that are monitored. Signal acquisition has to be noninvasive and as transparent as possible to the monitored subject. Hence, a limited number of biosignals are available, which include skin temperature, galvanic skin response, electrocardiogram (ECG), pupil diameter, and respiration.
- The choice of the type of emotion that the algorithm will predict. Emotions are controversially defined, but a common set of basic emotion labels is: sadness, happiness, fear, anger, surprise, and disgust, as defined by Ekman [6].
- The choice of a safe method to induce these types of emotions. The most used methods are video clips that contain a scene corresponding to the desired emotion, audio clips composed of music or sounds

- *G.E. Sakr and I.H. Elhajj are with the Department of Electrical and Computer Engineering, American University of Beirut, Beirut, Lebanon. E-mail: {ges07, ie05}@aub.edu.lb.*
- *H. Abou-Saad Huijer is with the Hariri School of Nursing, American University of Beirut, Beirut, Lebanon. E-mail: huda.huijer@aub.edu.lb.*

Manuscript received 17 Dec. 2009; revised 6 May 2010; accepted 21 June 2010; published online 19 July 2010.
Recommended for acceptance by S. Lee.
For information on obtaining reprints of this article, please send e-mail to: taffc@computer.org, and reference IEEECS Log Number TAFFC-2009-12-0010.
Digital Object Identifier no. 10.1109/T-AFFC.2010.2.

that correspond to the desired induced emotion, or a mix of both methods.

- The choice of the features that are extracted from the biosignals as well as the recognition algorithm that is used to detect emotion. In general, the features extracted are from the ECG, which commonly are the total energy, the RMS value, and the interbeat interval. The recognition algorithms include K-means or C-mean clustering, regression algorithms such as canonical regression, neural networks, support vector regression, and others.

ECG, skin temperature, skin conductance, and respiration have been used to extract 22 features for emotion recognition [7]. The induction of fear, joy, and neutrality was done using video clip stimuli. The authors used canonical correlation to achieve an 85.3 percent correct classification ratio. This method can be improved by reducing the number of features extracted from the ECG and trying to omit respiration from the monitored signals, which can be uncomfortable for the subject. Another work developed emotion recognition using ECG, skin temperature, and electrodermal activity, another name for GSR. Four different emotions were induced, sadness, anger, stress, and surprise, by reading a story with both the voice and the story reflecting the emotion induced. Then support vector machine for regression was used for classification, and 67.5 percent of correct-classification ratio was obtained [8]. The method can be improved by using multiclass SVM or by using one of its variants. Other work proposed detecting human emotion using only the ECG. From the ECG, three features were extracted: energy recoursing, energy efficiency, and root mean square. Four emotions were induced using audiovisual methods: happiness, disgust, surprise, and fear. Classification was done using unsupervised fuzzy C-mean clustering, where db4 wavelet transform was used for feature extraction. No accuracy was reported, but the graphs show good accuracy of clustering [9]. Fisher projection was used with sequential floating forward search to achieve an accuracy of 81 percent in detecting eight classes of emotion, including neutral. This study has proven the feasibility of emotion detection and that the features extracted for all of the emotions on a specific day are clustered more tightly than the features of a specific emotion during multiple days [4].

A high percentage of the agitation management techniques requires efficient and accurate agitation detection. Although not all for patients with dementia, many different agitation detection techniques were previously developed. Some techniques use hidden Markov models to estimate the inputs to an SVM, using an outside camera to observe the movement of the patient while sleeping [10]. The limitations of this approach are privacy concerns and the inability to monitor the patient all day. Another limitation is in the Markov model used. It is known that any model is called a Markov model if: knowing the present, the future is independent of the past. But for agitation detection, knowing the present is not always enough to predict the future, and historical behavior must be known. Some other techniques of frustration detection use facial expression, head movement, and eye movement. A dynamic Bayesian network model was used to integrate all of these features as well as GSR, RTD, and Blood Volume Pressure (BVP) [11]. The user must be sitting in front of a computer and holding the mouse to be able to measure all of the inputs, which is a limitation for elderly subjects and cannot be used clinically. A theoretical limitation to this method resides in the fact that using a Bayesian model means that the underlying probability densities between the output of the system and its inputs are known. But this probability density is, in most cases, not known and cannot even be estimated; hence, the need for a learning algorithm that does not need the underlying probability densities like SVM. Some studies done on the acceleration of the wrists, ankles, and waist found that agitated people had sudden movements of the wrists and ankles [12].

The most relevant study that was done so far was based on the measurement of BVP, GSR, ST, and the pupil diameter. From the heart rate, they extracted the Inter Beat Interval (IBI), and then studied it in the frequency domain because the ratio of the low frequency component over the high frequency component could be an indicator of stress [13]. Eleven features were extracted from these sensors and then fed into an SVM. The elicitation of stress was done using the Stroop test [14]. The SVM was tested using the cross-validation technique, and results showed 90.1 percent accuracy [13]. The limitation of this technique is the use of 11 features that are fed into the SVM, which is computationally demanding for a micro controller implementation. Moreover, the user has to sit in front of a computer to be able to capture the pupil diameter.

Previous work done by our team has shown that subject-independent agitation detection was possible by monitoring three vital signs: HR, GSR, and RTD, and, using SVM, it was possible to detect the agitation state of a subject even if the device was not trained on that specific subject. This was achieved with an accuracy that reached 84 percent [15]. The limitation of the method was the issue of the "gray zone." The gray zone is the transitional phase where the subject is passing from one state to another. In this area falls more than 95 percent of the detection errors of the method. Another work done by our team has also shown that using SVM with a hierarchical architecture that takes into consideration the distance between the two classes and hence giving the SVM a "zooming" ability gave high accuracy, reported at 96.16 percent. However, a bug was discovered in the code where points, not being mapped into any class, were being considered as correctly classified instead of wrongly classified. With this bug corrected the accuracy that should have been reported is 95.1 percent, which is still higher than what was previously reported in the literature. Either way, in this case the errors that were falling in the "gray zone" were also of high percentage [16]. The main question remaining is how to define the point between not agitated and agitated, and what the true ground truth is.

"Is it possible to define a transitional phase that detects the shifting of a subject from the relaxed state to the agitated state?" The two previous works published by our team were not able to detect agitation transition and did not introduce a confidence measure [15], [16]. The multilevel SVM introduced was based on the distance between the points and the centroid of the training class, and did not include any measure of confidence [16].

The aim of this paper is to show that using SVM, it is possible to define and detect the transitional state of a patient. In order to achieve this aim, a new confidence measure on the decision of a support vector machine is introduced. This confidence measure is used to modify a single 2-class SVM classifier into a single 3-class SVM classifier. To the best of our knowledge, this has not been developed in the literature. This new 3-class classifier is the basis of agitation transition detection. This new confidence measure is also used to modify a single 2-class SVM classifier into a multilevel 2-class SVM classifier architecture. This new architecture is used for agitation detection and yields, using the same training set, higher accuracy than single SVM 2-class classifier.

3 Proposed Detection Algorithm

In this section, we present the derivation of the detection algorithm as well as its implementation in two different architectures.

3.1 Feature Selection

The Central Nervous System (CNS) has two main sections: the somatic part, which is responsible for all of our voluntary movements, and the autonomic part, which controls all of the nonvoluntary movements, specifically those of the heart. The autonomic system consists of two main sections: the Sympathetic Nervous System (SNS), which is concerned with all emergency cases that we encounter in our lives, and the Parasympathetic Nervous System (PNS), which is responsible for our relaxed state. Finding the activity of the PNS and SNS leads to knowledge about the stress of the subject. High PNS activity indicates a relaxed patient, while high SNS activity corresponds to agitation.

Stress studies showed a high correlation between Heart Rate Variability (HRV), PNS, and SNS [17]. In most clinical applications, HRV is analyzed in the time domain and in the frequency domain. An important measure of the heart is the R-R interval, from which it is possible to extract the IBI. Indeed, the IBI can be extracted by taking any point as the reference point, but, since the R peak has the highest amplitude, it is immune to possible sources of noise [18], [19]. In the frequency domain, studies have shown that power distribution has four main spectral components: High Frequency (HF), Low Frequency (LF), Very Low Frequency (VLF), and Ultra Low Frequency [20]. They have also shown a major correlation between HF and PNS and between LF and SNS [17]. LF is calculated from the IBI, which is extracted from the normal to normal interval. Therefore, IBI is used as one of the features for agitation detection.

In order to improve detection, skin conductance and skin temperature were also monitored. These two signs were chosen because their measurement is noninvasive and they are in a direct relation to stress status. When subjects are stressed, blood vessels are contracted, which leads to a drop in the skin temperature. Studies have shown that the skin temperature is at its lowest when stress is at its maximum [21]. Also, the galvanic skin response is an indicator of stress level. Studies have shown that when the stress level goes up, the moisture of the body also goes up, which leads to a decrease in the resistance of the skin. In order to validate the necessity of using all of the above-mentioned features, a study using Principal Component Analysis (PCA) was conducted [22]. PCA is a well-known method that reduces a number of possibly correlated features to a smaller number using their covariance matrix: A feature with high variance is most likely to have more information and a feature that is independent of the other features is also of high importance because it carries new information. PCA is performed as follows: First, the data are centered by subtracting from each feature its mean, then the covariance matrix is computed, and finally, an eigenvalue/eigenvector decomposition is performed on the covariance matrix. The eigenvalues are used to classify the features in order of variability in each feature. The feature having the highest value is the one that has the most variability and, hence, contains most of the information. For agitation detection, the results for the three eigenvalues are the following: $\lambda_1 = 0.66$, $\lambda_2 = 0.28$, and $\lambda_3 = 0.11$. The values of the eigenvalues are comparable and, hence, taking away one feature will induce a significant loss of information; thus, the three features are used.

The challenge for a prediction systems is how to combine the different indicators in order to make a decision, and how to predict a large number of unseen patterns from a few known ones. The prediction has to be accurate, consistent, and computationally effective. This paper deals with the prediction problem; it presents an algorithm for detecting agitated and nonagitated states as well as an additional output that will predict the transitional phase between the two states. This algorithm is based on Support Vector Machines, which generalizes well, and on the new confidence measure introduced in this paper.

3.2 Support Vector Machines

In the simplest form, SVM uses a linear hyperplane to create a classifier with a maximal margin [23]. In other cases, where the data are not linearly separable, the data are mapped into a higher dimension feature space. This task is achieved using various nonlinear mapping functions: polynomial, sigmoid, and Radial Basis Functions (RBF) such as Gaussian RBF. In the higher dimension feature space, the SVM algorithm separates the data using a linear hyperplane. Unlike other techniques, probability model and probability density functions do not need to be known a priori. This is very important for generalization purposes as, in practical situations, there is not enough information about the underlying probability laws and distributions between the inputs and the outputs. Since SVM has been recording state-of-the-art accuracies in many fields, and, since it has an excellent generalization ability, it is used in the course of this paper.

What follows is an introduction to the theory of SVM and the general equation of the hyperplane that will separate the two classes. In the case of linearly separable data, the approach is to find among all the separating hyperplanes, the one that maximizes the margin. Clearly, any other hyperplane will have a greater expected risk than this hyperplane.

During the learning stage, the machine uses the training data to find the parameters $\mathbf{w} = [w_1 w_2 \dots w_n]^T$ and b of a decision function $d(\mathbf{x}, \mathbf{w}, b)$ given by:

$$d(\mathbf{x}, \mathbf{w}, b) = \mathbf{w}^T\mathbf{x} + b = \sum_{i=1}^{n} w_i x_i + b. \quad (1)$$

The separating hyperplane follows $d(\mathbf{x}, \mathbf{w}, b) = 0$. In the testing phase, an unseen vector x will produce an output y according to the following indicator function:

$$y = sign(d(\mathbf{x}, \mathbf{w}, b)). \quad (2)$$

In other words, the decision rule is: If $d(\mathbf{x}, \mathbf{w}, b) > 0$, then $\mathbf{x}$ belongs to class 1, and if $d(\mathbf{x}, \mathbf{w}, b) < 0$, then $\mathbf{x}$ belongs to class 2.

The weight vector and the bias are obtained by minimizing the following equation:

$$L_d(\alpha) = 0.5\alpha^T H\alpha - f^T\alpha, \quad (3)$$

subject to the following constraints:

$$y^T\alpha = 0,$$
$$\alpha \geq 0,$$

where H denotes the Hessian matrix given by: $H = y_i y_j (x_i x_j)$ and f is the unity vector $f = [1, 1 \ldots 1]^T$. Having the solutions α_{0i} of the dual optimization problem will be sufficient to determine the weight vector and the bias using the following equations:

$$\mathbf{w} = \sum_{i=1}^{l} \alpha_{0i} y_i x_i, \quad (4)$$

$$b = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{y_i} - x_i^T\mathbf{w}\right), \quad (5)$$

where N represents the number of support vectors.

The linear classifier presented above has limited capabilities since it is only used with linearly separable data, while, in most practical applications, data are random and are not linearly separable. The nonlinear data have to be mapped to a new feature space of higher dimension, using a suitable mapping function, $\Phi(x)$, which is of very high dimension, potentially infinite. Fortunately, in all of the equations, this function appears only in the form of a dot product.

From the theory of reproducing kernel Hilbert spaces [24], which is beyond the scope of this paper, a kernel function is defined to be:

$$K(x_i, x_j) = \Phi(x_i)^T\Phi(x_j). \quad (6)$$

By replacing the dot product $x_i \cdot x_j$ by $K(x_i, x_j)$ in all of the previous equations, the nonlinear hyperplane is determined as:

$$d(\mathbf{x}) = \sum_{i=1}^{l} y_i \alpha_i K(x_i, \mathbf{x}). \quad (7)$$

This remarkable characteristic of the kernel transformation gives the ability for SVM to operate on multidimensional data without affecting the processing time. Indeed, in the linear case, the processing time is roughly the time needed to invert the Hessian matrix, which is of $\mathcal{O}(n^3)$, where n is the number of training points. Since the transformation from the linear to the nonlinear case is performed by the simple kernel transformation, the dimension of the Hessian matrix is not changed, and hence, the processing time is the same; thus its applicability and high performance in multidimensional data.

The solution of (3) yields the hard margin classifier. In general, it is useful to use a soft margin classifier to preserve the smoothness of the hyperplane and prevent α_i from tending to infinity. This classifier is obtained using the same minimization process by adding one more constraint to (3). The constraint is: $0 \leq \alpha_i \leq C$, where C is defined by the user. If C tends to infinity, the soft margin classifier tends toward the hard margin. In what follows, we define and discuss the natural confidence measure.

3.3 Natural Confidence Measure

The significance of the confidence measure extends beyond the accuracy of a classification problem. We live in a world where we are constantly making decisions and estimating our confidence in these decisions. For example, a reviewer has to make a decision about accepting or rejecting a certain paper and has to fill in the confidence in his decision. The ability of humans to give decisions as well as confidence measures in these decisions suggests that a confidence measure is useful in building classifiers.

Intuitively, a natural confidence measure in the decision $d(\mathbf{x})$ of an SVM would be inversely proportional to the distance of $\mathbf{x}$ to the hyperplane $\mathcal{H}$. As $\mathbf{x}$ gets closer to the hyperplane, the confidence in the decision decreases. Formally, the confidence could be defined as:

Definition 1. *Let $\mathbf{x}$ be an unseen vector, then the confidence C in the decision is given by:*

$$C = \frac{\gamma}{dist(x, \mathcal{H})}, \quad (8)$$

where $dist(x, \mathcal{H})$ is the distance of vector $\mathbf{x}$ to the hyperplane $\mathcal{H}$ and $\gamma \in \mathbb{R}^{*+}$.

If $\mathbf{y}$ is a vector on the hyperplane, then $dist(x, \mathcal{H})$ is obtained by minimizing the following equation:

$$dist(x, \mathcal{H})^2 = \min(x - y)^T(x - y), \quad (9)$$

subject to the constraint $d(y) = 0$, where d is the equation of the hyperplane given by (7). The global minimum for this minimization problem is not guaranteed to be found because of the complex nature of the decision function (7) as well as the lack of guarantee that $d(y) = 0$ is a continuous function.

Definition 1 does not capture the confidence of the decision accurately in all cases. Confidence should also be a function of the distance of the point to the training set. It is intuitive that if a point, z, is far from the hyperplane but also far from the training set, then it should have very low confidence. Four cases can be thought of:

- A point far from the plane and far from the training points should have very low confidence.
- A point close to the hyperplane and far from the training points should have low confidence.
- A point close to the hyperplane and close to the training set should have medium confidence.
- A point away from the hyperplane and close to the training set should have the highest confidence.

But, the challenge that arises is how to define the distance to the training points. It can be defined as the distance to their centroid or the minimal distance to the training points or the maximal distance to the training points. Ideally, one can imagine a two-dimensional classification problem, where all of the points to be classified lie on a plane. The decision algorithm with a confidence measure should be able to map one class above the plane and the other class under the plane. And the points left on the plane should be undecided (no confidence). However, a traditional classifier just groups one class in a region, with everything outside this region classified as the other class, and no measure of confidence is provided.

3.4 Proposed Confidence Measure

The proposed method is based on a dimension proposed by Vapnik and Chervonenkis, which was named after them: the VC dimension. By definition, the VC dimension is the capacity of the learning algorithm to shatter points in the input space [25]. Formally, it is the cardinality of the largest set of points that an algorithm can shatter. The importance of the VC dimension is that it appears explicitly in the bound on the total error of an algorithm. The total error of the learning machine is the sum of the training error (empirical error) and the testing error (generalization error):

$$\varepsilon = \varepsilon_{emp} + \varepsilon_g, \tag{10}$$

where ε_{emp} is the training error and ε_g is the generalization error [25]. ε_{emp} can be made arbitrarily small by choosing a machine with a VC dimension at least equal to the number of training points. In order to decrease the training error, one has to increase the VC dimension. If a machine can split all of the training points without any errors, it will have $\varepsilon_{emp} = 0$. Vapnik has established a bound on the testing error given by:

$$\varepsilon_g < \sqrt{\frac{VC[ln(\frac{2l}{VC}) + 1 - ln(\frac{\eta}{4})]}{l}}, \tag{11}$$

where l is the number of training data, VC is the VC dimension (referred to as h in some references), and $1 - \eta$ is the probability for which this last equation holds. This inequality shows that the error is bounded by an increasing function of the VC dimension, and thus a trade-off should be made between the empirical error and the generalization error. Although it is extremely difficult, and sometimes impossible, to compute the VC dimension of a certain algorithm, a bound on the VC dimension has been established and will be very useful in building the confidence measure. Vapnik states that a bound on the VC dimension is given by:

$$VC < \|w\|^2 D^2, \tag{12}$$

where D is the minimum radius of the sphere that contains all the training points and $\|w\|$ is the norm of the weight vector that SVM is minimizing. $\|w\|^2$ is given by [25]:

$$\|w\|^2 = \sum_{i=1}^{n} \alpha_i. \tag{13}$$

This bound is important in two ways: It is easy to compute, and Burges has shown that the true VC is closely related to this bound. In particular, he showed that, most of the time, the true minimum of the VC dimension is obtained when this bound is minimal [26].

Notation. *The following notations will be used in the definitions and propositions that follow:*

- $X = \{x_1, x_2, \ldots x_n\}$ *is the set of training vectors, where* $x_i \in \mathbb{R}^m$ *and* $Y = \{y_1, y_2, \ldots y_n\}$ *is their corresponding class, where* $y_i \in [-1, 1]$.
- $S(X, Y)$ *is any learning algorithm trained by training vectors X with class Y, and where all of the parameters of S are set.*
- D *is the radius of the smallest sphere that englobes all of the training points and* w *is the weight vector of the hyperplane.*
- z *is an unlabeled vector and* $d(z) \in [-1, 1]$ *is its prediction given by (7).*
- X_z *is the set of training points* X *to which* z *is appended,* $X_z = \{X, z\}$.
- Y_{-1} *is the set of labels* Y *to which* (-1) *is appended:* $Y_{-1} = \{Y, -1\}$, *and* Y_1 *is the set of labels* Y *to which* (1) *is appended:* $Y_1 = \{Y, 1\}$.

Based on the notation presented, X_z represents a new training set that contains X to which z is added. Y_{-1} is the set of labels formed by Y to which (-1) is added. This means that we are assuming that z is a training point having the label (-1), while Y_1 assumes that z has the label (1). It is now necessary to define a quantity that estimates the confidence in each of the assumptions (z belongs to class 1 or class -1).

Note that:

- To every $S(X, Y)$ corresponds a unique couple (D, w). This can be justified by the fact that D is unique and depends on the training points X, and w is a single valued function of S, X, and Y.
- Since the VC dimension is closely related to its upper bound, its upper bound $VC^{max} = D^2\|w\|^2$ will be used.
- As a consequence of the uniqueness of the couple (D, w) to every $S(X, Y)$, we will denote the VC dimension of $S(X, Y)$ by $VC_S^{max}(X, Y)$.

Definition 2. *Given* $S(X, Y)$ *and an unseen vector z, the measure of confidence on the decision* $d(z)$ *by* $S(X, Y)$ *is given by:*

$$C(z) = d(z)\big[VC_S^{max}(X_z, Y_{-1}) - VC_S^{max}(X_z, Y_1)\big]. \tag{14}$$

Note that $d(z)$ is the decision taken by $S(X, Y)$, while $S(X_z, Y_{-1})$ and $S(X_z, Y_1)$ are never used to take decisions, but are used only to measure their VC_S^{max}.

Equation (14) is a general measure of confidence, irrespective of the learning algorithm S. In the rest of this paper, the study is restricted to Support Vector Machines.

Proposition 1. *If* z *is on the hyperplane generated by* $S(X, Y)$ *(i.e.,* $d(z) = 0$*), then:*

1. $C(z) = 0$,
2. $VC_S^{max}(X_z, Y_{-1}) = VC_S^{max}(X_z, Y_1)$.

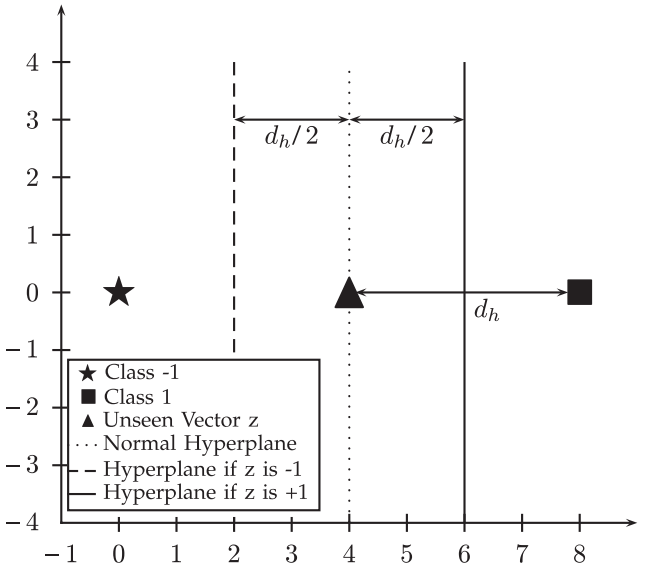

Fig. 1. Hyperplanes using $S(X,Y)$, $S(X_z,Y_1)$, and $S(X_z,Y_{-1})$.

Proof. 1) $C(z)=0$ is a direct result of the definition by noticing that using the soft margin classifier $\|w\|$ is always finite and thus VC_S^{max} is also finite.

2) Recall that $VC_S^{max}=D^2\|w\|^2$ and notice that, in $S(X_z,Y_{-1})$ and $S(X_z,Y_1)$, the radius of the smallest sphere that contains all of the training points (D) is the same. So it is sufficient to prove that $\|w\|$ is also the same in both cases. Note that the distance between a support vector and the hyperplane is given by [23]:

$$d_h=\frac{1}{\|w\|}. \tag{15}$$

If z is on the hyperplane, then the new hyperplane will be a simple translation of the old hyperplane by a distance of $\frac{d_h}{2}$ toward the other class, i.e., in the case of $S(X_z,Y_{-1})$, the hyperplane moves toward class (1), while, in the case of $S(X_z,Y_1)$, the hyperplane moves toward class (-1) by the same distance as illustrated in Fig. 1. Since the distance is the same in both cases, we can deduce from (15) that $\|w\|$ is the same in both cases and, since D is also the same, then $VC_S^{max}(X_z,Y_{-1})=VC_S^{max}(X_z,Y_1)$. □

Note that the only proof needed is for the linearly separable case because, for the nonlinear case, SVM will map the data into a higher dimensional feature space $\mathcal{F}$, in which the training data are linearly separable. This is a result of the fact that the training error in support vector machines can always be made zero.

As stated in the discussion of natural confidence, an important property of the confidence measure is that C should satisfy the four cases of confidence. This property is achieved by careful choice of the kernel function, as shown in Proposition 2.

Proposition 2. *Let z be a vector far from any training point such that: $\|z-x_i\|\to\infty, \forall x_i\in X$.*

If $K(x_i,x_j)$ is such that $0\le K(x_i,x_j)\le 1$ and $\lim_{\|x_i-x_j\|\to\infty}K(x_i,x_j)=0$, then $C(z)\to 0$.

Proof. Two conditions need to be satisfied for $C(z)\to 0$: $d(z)\to 0$ and $VC_S^{max}<\infty$. By using the soft margin classifier, it is guaranteed that: $\alpha_i<\infty,\forall i$. By applying (7), it is clear that every term in the finite sum tends to zero because $K(x,x_i)\to 0$, and hence $d(z)$ tends to 0.

To prove that $VC_S^{max}<\infty$, it is only necessary to prove that D^2 is finite because it is already known by (13) that $\|w\|^2$ is finite.

By using a kernel function, every point x_i is mapped to a new space (the feature space) using the mapping function $\phi(x_i)$. The distance between any two points in the feature space is given by:

$$D_f^2(x_i,x_j)=(\phi(x_i)-\phi(x_j))^T(\phi(x_i)-\phi(x_j)).$$

Expanding and using (6):

$$D_f^2(x_i,x_j)=K(x_i,x_i)+K(x_j,x_j)-2K(x_i,x_j).$$

But, $0\le K(x_i,x_j)\le 1$ implies that $D_f^2(x_i,x_j)$ is finite $\forall x_i,x_j$. Since D^2 is the square of the distance between the center of the sphere and the training points that lie on the boundary, then it is also bounded and, hence, VC_S^{max} is bounded, which implies that $C(z)\to 0$. □

Proposition 2 presents a sufficient condition on the kernel function. One kernel that satisfies this condition is the Radial Basis Function (RBF) kernel:

$$K(x,x_i)=e^{-(x-x_i)^T\Sigma^{-1}(x-x_i)},$$

where Σ^{-1} is a covariance matrix. This type of kernel function satisfies the condition of Proposition 2 because the covariance matrix is always a positive semidefinite matrix. Indeed, a positive semidefinite matrix A satisfies $x^tAx\ge 0\forall x$, and

$$\begin{aligned}
&\text{if} && \|x-x_i\|^2\to\infty,\\
&\text{then} && (x-x_i)^T\Sigma^{-1}(x-x_i)\to+\infty,\\
&\text{thus} && K(x,x_i)\to 0,\\
&\text{and} && -(x-x_i)^T\Sigma^{-1}(x-x_i)\le 0,\\
&\text{thus} && 0\le K(x,x_i)\le 1.
\end{aligned}$$

The Gaussian kernel is the case where the covariance matrix is a diagonal matrix with identical value σ^2:

$$K(x,x_i)=e^{-\frac{1}{2\sigma^2}(x-x_i)^T(x-x_i)},$$

where σ is the width of the gaussian function. This kernel also satisfies Proposition 2:

$$\begin{aligned}
&(x-x_i)^T(x-x_i)=\|x-x_i\|^2.\\
&\qquad\text{if} \quad \|x-x_i\|^2\to\infty,\\
&\qquad\text{then} \quad K(x,x_i)\to 0,\\
&\qquad\text{and} \quad -\frac{1}{2\sigma^2}(x-x_i)^T(x-x_i)<0,\\
&\qquad\text{thus,} \quad 0\le K(x,x_i)\le 1.
\end{aligned}$$

The choice of the Gaussian kernel function is not limiting because this kernel function is used in a large percentage of applications. It is a very popular kernel function because a bias term is not needed in the decision function, which leads to an easier quadratic optimization problem. The second reason is that the number of parameters that influence the complexity of the model in the Gaussian function is much less than the polynomial function. Finally,

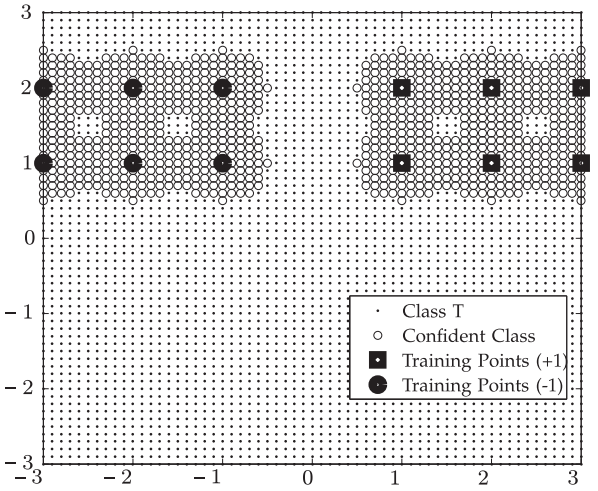

Fig. 2. 3-classes SVM.

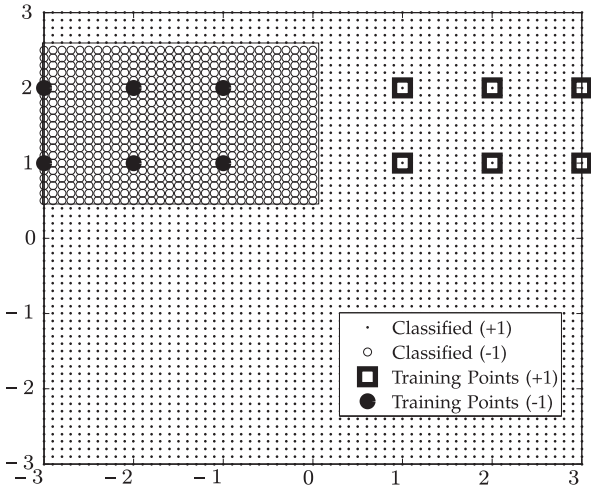

Fig. 3. Normal SVM hyperplane.

the Gaussian kernel has less numerical complexity because it is bounded between 0 and 1, while the polynomial function can tend to infinity.

3.5 Proposed Architectures

The proposed classifiers are based on the properties of $C(z)$. Proposition 1 shows that the limit of the confidence is given by $C(z) = 0$. In addition, $C(z) < 0$ implies that the decision was taken in favor of one class, while z looks more like the other class. Indeed, suppose $d(z) < 0$, thus z was classified as class (-1). But if $C(z) < 0$ implies that $VC_S^{max}(X_z, Y_{-1}) - VC_S^{max}(X_z, Y_1) > 0$, which means that if z was a training point of class (-1), it would have generated a classifier that has a higher VC dimension than the one that would have been trained with z as class (1). The error bound given by (11) suggests that the first classifier (the one that has the higher VC) will have a higher probability of error, which also suggests that the point should rather be in class (1). The same reasoning holds if, initially, z was classified as class (1) and $C(z) < 0$, then it should have been of class (-1). $C(z)$ could be used in several ways; Here, two methods that are used in agitation detection are presented.

3.5.1 Agitation Transition Detection

Section 3.2 has shown that an SVM classifier takes an input x and classifies it as class (1) or class (-1). Traditionally, to obtain a 3-class SVM one has to cascade two SVMs: The first one has to decide if the input is of class (1) or not class (1). If not class (1), then it goes into the second SVM that is trained to decide between class (2) and class (3) [25]. Based on $C(z)$, it is possible to define the three classes using only one SVM.

Definition 3. *Let z be an input of unknown class, then:*

$$Class = \begin{cases} -1, & d(z) < 0 \text{ and } C(z) > 0, \\ 1, & d(z) > 0 \text{ and } C(z) > 0, \\ T, & C(z) \leq 0, \end{cases} \tag{16}$$

where "T" is the transitional phase of agitation.

This 3-class classifier is used for agitation transition detection. If the output of the system is (-1), then the subject is not agitated. If the output of the system is $(+1)$, then the subject is agitated. If the output is (T), then the subject is in the transitional phase: either was calm and is getting agitated or was agitated and is calming down. The agitation transition detection is possible with the addition of the confidence measure to the normal 2-class classifier. The focus will be on the case where the subject is getting agitated because the ultimate goal is early prediction of agitation. To illustrate this classifier, six training points of each class are chosen to be linearly separable (Fig. 2). Class (-1) is represented by the big filled dots and class (1) by the large filled squares. Fig. 2 shows the dark area around the training points of each class. It consists of empty circles that correspond to the confident area: $(C(z) > 0)$ and are classified correctly ($+1$ near the large filled squares and -1 near the big filled dots). It also shows a light area that consists of light dots that correspond to the nonconfident points $(C(z) < 0)$. Fig. 2 shows that, according to the discussion of Section 3.3, a classifier, along with a confidence measure, should be able to map confident points above and under the plane, while leaving nonconfident points on the plane is achieved. It could be visualized as if the dark area corresponding to the points near class (1) are above the horizontal plane, while the points next to class (-1) are under the plane and the remaining points are on the plane, while the hyperplane generated by a traditional SVM, and shown in Fig. 3, shows that everything that falls inside the closed area is of class (1), while the ones outside the closed area are of class (-1), where many correspond to classification errors.

Instead of being used to define a transitional phase, $C(z)$ could be used to enhance the two-class classification accuracy, as discussed in the next section.

3.5.2 Multilevel SVM

The multilevel SVM is a two-class classifier based on cascading two SVMs. The first SVM (SVM-1) is trained by "easily classifiable" points, while the second one (SVM-2) is trained by not "easily classifiable" points. An "easily classifiable" point is defined as follows:

Definition 4. *Let X be the set of training points and $x \in X$ is one specific training point. x is said to be easily classifiable iff $C(x) > 0$.*

Note that the confidence is computed by SVM-1; the output of this architecture is defined as follows:

Definition 5. *Let z be an input of unknown class, then:*

$$Class = \begin{cases} -1, & d_1(z) < 0 \text{ and } C_1(z) > 0, \\ & \text{or } d_2(z) < 0 \text{ and } C_1(z) < 0, \\ 1, & d_1(z) > 0 \text{ and } C_1(z) > 0, \\ & \text{or } d_2(z) > 0 \text{ and } C_1(z) < 0. \end{cases} \quad (17)$$

An unknown point z is classified by SVM-1 and a measure of confidence is then generated. If SVM-1 is confident of its decision, then this decision will be the final one. If SVM-1 is not confident, then the point is sent to SVM-2, and its decision will be the final decision.

Note that this architecture is recursive and could be extended to SVM-n, and the unknown point will go deep in the architecture until the decision made by SVM-j is confident and then its decision will be the final one. In the agitation detection application, two SVMs were enough to classify all the points.

The next section gives a brief description of the experimental setup of data collection as well as the results obtained by the methods described in this section.

4 Experiment Design and Results

The experiment objective was to carry out a quantitative study designed to evaluate the onset of agitation, using healthy subjects. One effective way of inducing stress safely into healthy subjects is the Stroop Color-Word Interference Test. The Stroop test requires subjects to say out loud the color of words, spelling out color names that do not match. It has been shown that this test induces anxiety symptoms [14].

The Stroop Color-Word Interference Test in its classical version has been widely used as a psychological or cognitive stressor that can safely induce controlled limited stress in subjects. Previous research has indicated that by adding task pacing to the Stroop test, physiological responses intensify. Three sets of PowerPoint slides were used to conduct the Stroop test while the physiological parameters were monitored. The first set is called the color blocks (show a block of color only), the second set is called congruent word slides (word and color matches), and the third set is called incongruent word slides (words do not match color presented). The test included 60 randomly set color blocks (1 min), where subjects are to name out loud the color they see, 60 randomly set congruent word slides (1 min) where subjects are to read out loud the word they see, and finally, 120 randomly set incongruent word slides (2 min) where subjects are to name the color of the word they see (not read the word). Each slide is set to show for one second. If the subjects missed a slide, they were asked to move on to the next one. The overall objective of the subject was to get as many correct answers as they could. Subjects were asked to complete the Trait Scale State-Trait Anxiety Inventory (T-STAI) before and after the test. The two STAI filled up by a subject allowed us to validate their state.

As discussed previously, the features used are the skin temperature, the galvanic skin response, and the HRV. From the heart rate, the Inter-Beat Interval (IBI) was extracted. HRV has been used extensively in reflecting the way the central nervous system works, specifically the sympathetic side. Although the amplitude of the heart beat is also correlated to the stress level, it was not used as an input to the SVM in order to reduce calculations and power consumption, which is imperative for the portability of the device.

4.1 Data Collection and Processing

To measure and record the physical features, the following sensors were used: a polar exercise heart rate monitor from Vernier, a 1,000 ohms platinum resistance temperature detector (RTD) from Omega, and electrodes that wrap around the fingers for monitoring galvanic skin response. The RTD sensor changes its resistance with the skin temperature of the subject. The change in resistance is converted into temperature change by using the Callendar-Van Dusen equation:

$$R_t = R_0 + R_0\alpha\left[t - \delta\left(\frac{t}{100} - 1\right)\left(\frac{t}{100}\right) - \beta\left(\frac{t}{100} - 1\right)\left(\frac{t}{100}\right)^3\right]. \quad (18)$$

Since the measured temperature is always above 0, only the Callendar coefficient δ is used, while the Van Dusen coefficient $\beta = 0$ is for positive temperatures.

The experimental procedure is as follows: The subject places the sensors around his body. He then undergoes the Trait scale State-Trait Anxiety Inventory (T-STAI) [27]. The trait-anxiety scale is one of two subscales of the full form STAI developed by Spielberger to measure anxiety in adults. It is one of the most frequently used measures of anxiety in applied psychology research and has been shown to be a reliable and sensitive measure of anxiety. Subjects were asked to fill the T-STAI before and after the Stroop test. When undergoing the Stroop test, all of the signals are being recorded on the same machine where the Stroop test is running [14]. A sample is generated per second during the 4-minute test, which yields a total number of 240 samples per subject. One important aspect of (and probably the biggest issue relating to) biosignals is the baseline measurement for each person. The proposed approach to solve this problem is by normalizing the data, which are simple but proved effective in allowing subject-independent agitation detection. Clearly, this normalization requires the statistics (mean, standard deviation) of the features used. However, collecting the data to obtain the statistics is much easier than establishing an agitation baseline for a subject. Therefore, this approach results in the most subject-independent agitation detection we are aware of.

If RTD_i is the temperature at a certain point i and μ_r is the mean of the skin temperature defined by $\mu_r = \frac{1}{n}\sum_{i=1}^{n} RTD_i$ with standard deviation σ_r, then the normalized value is:

$$RTD_{in} = \frac{RTD_i - \mu_r}{\sigma_r}.$$

If GSR_i is the galvanic skin response at a point i, and the mean of the GSR is given by $\mu_g = \frac{1}{n}\sum_{i=1}^{n} GSR_i$ with standard deviation σ_g, then the normalized value is:

$$GSR_{in} = \frac{GSR_i - \mu_g}{\sigma_g}.$$

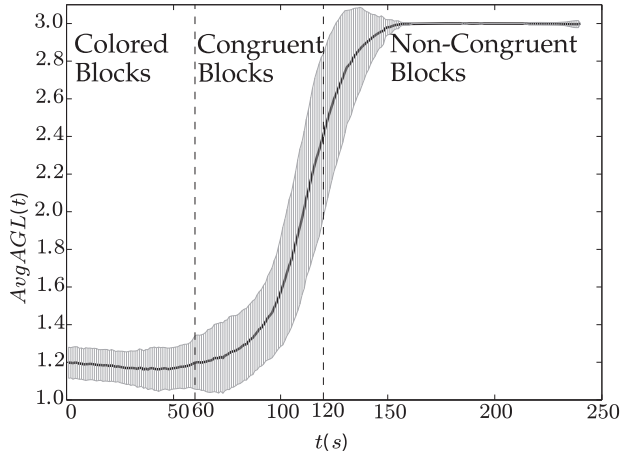

Fig. 4. $AvgAGL(t)$.

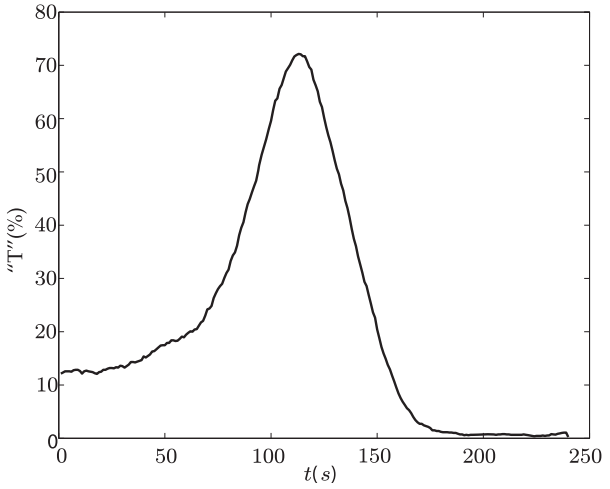

Fig. 5. Percentage of points classified in the transitional phase.

If IBI_i is the IBI at a point i and the mean of the IBI is given by $\mu_b = \frac{1}{n}\sum_{i=1}^{n} IBI_i$ with standard deviation σ_b, then the normalized value is:

$$IBI_{in} = \frac{IBI_i - \mu_b}{\sigma_b}.$$

After this normalization, it is possible to train the learning algorithm with a group of subjects and test it on subjects not belonging to that group.

To reduce the noise, and knowing that the correlation between two consecutive measurements in time of the above feature is very high, it is very useful to apply exponential decay on the normalized data. Although this is not an advanced filtering technique, the ultimate aim of this paper is to develop a low power portable device that the subject will carry and will alert caregivers whenever the patient is agitated. In addition, since the results were not very sensitive to the filtering method used, this type of primitive filtering was chosen. If X_n represents any of the above normalized features, then

$$X_{n_{i+1}} = \alpha X_{n_i} + (1-\alpha) X_{n_{i+1}}.$$

Choosing α to be close to 1 will make the data more dependent on the previous measures, and choosing α close to 0 will make the data less dependent on the previous measure. The value chosen for this experiment is $\alpha = 0.8$.

4.2 Transitional Agitation Detection Results

Data from subjects that demonstrated stress using the T-STAI analysis were used to evaluate the accuracy of the detection algorithms. Each patient had 240 samples (RTD, GSR, IBI) taken synchronously corresponding to the three Stroop test slides. In total, 58 subjects were tested. The agitation level (AGL) of a sample X of subject s taken at time t is defined as follows:

$$AGL(X_s(t)) = \begin{cases} 1, & \text{Non Agitated State,} \\ 2, & \text{Transitional State,} \\ 3, & \text{Agitated State.} \end{cases} \quad (19)$$

Also, define $AvgAGL(t)$ to be the expected value of the agitation level, over all subjects, at time t:

$$AvgAGL(t) = E_s[AGL(X_s(t))]. \quad (20)$$

The proposed algorithm is tested using the K-fold cross-validation technique. The 58 subjects are subdivided into groups of two to form 29 folds. In general, the cross-validation is carried out by taking 28 folds as the training set and the remaining fold as the validation set. The final result is the average of all K permutations. To illustrate the robustness of our architecture against overfitting, which, in general, is caused by the lack of training points, only one fold is taken as the training set and the remaining 28 folds are taken as the validation set. The expectation in (22) represents an ensemble average over all 28 permutations. In order to choose the best parameter $1/\sigma^2$ for the architecture, a subset of 12 subjects is considered. A 6-Fold cross-validation was conducted, where $1/\sigma^2$ was varied over a large range. The value of $1/\sigma^2$ that gave the best average performance over the 12 subjects was chosen in the larger 29-fold cross-validation procedure. It is worth noting that, in all cross validation procedures for traditional and multilevel SVM, no training points were used for testing. This is a standard procedure to avoid biasing the accuracy of the testing phase. Fig. 4 shows $AvgAGL(t)$, where the points corresponding to $t < 100s$ have an average agitation level of 1.2, while the points corresponding to $100s \leq t \leq 140$ have an average agitation level of 2.2, and the points corresponding to $t > 140s$ have an average of 2.99. The gray area corresponds to a standard deviation around the average. As expected, the subject is relaxed for the first 100 points then goes into a transitional phase of around 40 seconds, and then remains agitated for the remainder of the experiment. It is also notable that for the first 20 seconds, the agitation level is decreasing, which shows that some subjects were not totally relaxed at the beginning of the experiment. Note also that, although it is not clear from Fig. 4, a limited number of subjects exhibits a decrease in the agitation level during the last 10 samples. This could show that some subjects got familiar with the noncongruent segments and started relaxing before the end of the experiment. Fig. 5 shows the average percentage of subjects over all 29-fold permutations that were classified in the transitional phase at time "t." This figure demonstrates that indeed the points inside the interval $[100s, 140s]$ are in the majority of class 2, and hence their average is due to their correct classification rather than to the fact that they were equally misclassified between 1 and 3 and hence averaged to 2.2. A standard assumption in the

literature is that the subject would remain calm $(AGL = 1)$ during the first 120 samples (colored blocks and congruent blocks), and then get agitated $(AGL = 3)$ for the remaining 120 samples (noncongruent blocks). This assumption was used [13], [15], [16] as the definition of the ground truth to compute the accuracy of agitation detection of one SVM. To be consistent with the literature, this assumption will hold in the next section, where it is shown that the proposed multilevel SVM outperforms the single SVM.

4.3 Multilevel SVM Results

In this section, only two classes exist:

$$\begin{cases} -1, & 1 \le t \le 120, \\ 1, & 120 < t \le 240. \end{cases} \quad (21)$$

For this experiment, the same 58 subjects were considered. In order to test this architecture, the K-fold cross-validation technique is also used. The 58 subjects are subdivided into groups of two, which yields 29 folds. In order to show the robustness of this architecture against overfitting, only one fold is taken for training and the remaining folds are used for validation. From the training fold, a subset X of 80 triplets (ST, GSR, IBI) is chosen randomly. X is used to train a traditional single SVM using the gaussian kernel function as discussed earlier, and all the samples of the remaining 28 folds were used to test its accuracy. The same set X was split, using the method described in Section 3.5.2, into two sets X_1 and X_2. SVM-1 was trained using X_1 while SVM-2 was trained using X_2, and the samples from the remaining 28 folds were used in the testing procedure as also described in Section 3.5.2. The comparison of the performance of a single SVM with the performance of the multilevel SVM is fair because, in both cases, the same training set X is used. Note that for the normal SVM and the multilevel architecture, the parameter $1/\sigma^2$ has to be determined. For this purpose, a reduced set of 12 subjects was used. A 6-fold cross-validation was carried out on both normal and multilevel architectures over 50 different values of $1/\sigma^2$. The values of $1/\sigma^2$ that yielded the highest accuracies for both architectures are used for the 29-fold validation procedure. The average accuracy for the proposed architecture reached 91.4 percent, with a standard deviation of 1.71 percent, while the average accuracy for the normal SVM reached 90.9 percent, with a standard deviation of 1.94 percent. The best accuracy for the normal SVM was reached with $1/\sigma^2 = 50$, while the best accuracy for the proposed architecture was reached using $1/\sigma^2 = 2$. It was shown that the value of $1/\sigma^2$ is an indicator of the VC dimension: If a machine is trained using a higher value of $1/\sigma^2$ than another machine, then its VC dimension is also higher [23]. And since the generalization error is bounded by an increasing function of the VC dimension, as shown in (11), then our architecture is more likely to generalize better over unseen new data. The reported accuracy cannot be compared with accuracies reported in the literature because the algorithms were tested on different data sets and using different validation techniques.

Since this system is motivated by a medical application, there is an important trade-off between usability and detection. However, from a practical perspective, having an accuracy in the 90 percent range is acceptable as this system is intended to assist caregivers and not replace them completely. So it is expected that when there is an alert, a caregiver would conduct a more thorough diagnosis. It is the usability of the system for such an application that motivated the subject-independent method. Since it is very challenging to obtain a baseline by reducing the measurements required as described previously, the system application became more practical.

5 Conclusion

This paper presented a decision confidence measure and two new SVM architectures, which were applied to agitation detection and agitation transition detection. It was shown that, under no constraints on the subject and using three vital signs, the transitional phase was detected with high accuracy using this new confidence measure. It was also shown that, using the confidence measure, it is possible to build an architecture that yields higher accuracy than the traditional SVM, while using the same training set. An accuracy of 91.4 percent was achieved, in comparison with 90.9 percent for the traditional SVM.

Acknowledgments

The authors thank Dr. Cheryl Riley-Doucet and Dr. Debatosh Debnath from Oakland University for providing the data used in this research. This research was funded by the American University of Beirut University Research Board, Dar Al-Handassah (Shair & Partners) Research Fund, and the Rathman (Kadifa) Fund.

References

[1] "DSM-IV: Diagnostic and Statistical Manual of Mental Disorders," *Am. Psychiatric Assoc. Task Force on DSM-IV,* 1994.

[2] W. He, M. Sengupta, V. Velkoff, and K. DeBarros, "US Census Bureau, Current Population Reports, P23-209, 65+ in the United States: 2005," US Govt. Printing Office, 2005.

[3] A. Rosenblatt, "The Art of Managing Dementia in the Elderly," *Cleveland Clinic J. Medicine,* vol. 72, no. 3, p. 3, 2005.

[4] R. Picard, E. Vyzas, and J. Healey, "Toward Machine Emotional Intelligence: Analysis of Affective Physiological State," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 23, no. 10, pp. 1175-1191, Oct. 2001.

[5] W. Liao, W. Zhang, Z. Zhu, and Q. Ji, "A Real-Time Human Stress Monitoring System Using Dynamic Bayesian Network," *Proc. 2005 IEEE CS Conf. Computer Vision and Pattern Recognition,* p. 70, June 2005.

[6] P. Ekman, "Expression and the Nature of Emotion," *Approaches to Emotion,* vol. 3, pp. 319-343, Lawrence Erlbaum Assoc., 1984.

[7] L. Li and J. Chen, "Emotion Recognition Using Physiological Signals," *Advances in Artificial Reality and Tele-Existence,* pp. 437-446, Springer, 2006.

[8] K. Kim, S. Bang, and S. Kim, "Emotion Recognition System Using Short-Term Monitoring of Physiological Signals," *Medical and Biological Eng. and Computing,* vol. 42, no. 3, pp. 419-427, 2004.

[9] M. Murugappan, M. Rizon, R. Nagarajan, S. Yaacob, D. Hazry, and I. Zunaidi, "Time-Frequency Analysis of EEG Signals for Human Emotion Detection," *Proc. Fourth Kuala Lumpur Int'l Conf. Biomedical Eng.,* pp. 262-265, 2008.

[10] V. Fook, P. Thang, T. Htwe, Q. Qiang, A. Wai, M. Jayachandran, J. Biswas, and P. Yap, "Automated Recognition of Complex Agitation Behavior of Dementia Patients Using Video Camera," *Proc. Ninth Int'l Conf. e-Health Networking, Application and Services,* pp. 68-73, 2007.

[11] L. Wenhui, Z. Weihong, Z. Zhiwei, and Q. Ji, "A Real-Time Human Stress Monitoring System Using Dynamic Bayesian Network," *Proc. IEEE CS Conf. Computer Vision and Pattern Recognition,* June 2005.

[12] T. Tamura, T. Fujimoto, and T. Togawa, "Quantitative Assessment of Behavior in Dementia Patients by Continuous Physical Activity Monitoring," *Proc. 19th Ann. Int'l Conf. IEEE Eng. in Medicine and Biology Soc.*, vol. 3, pp. 999-1002, 1997.
[13] J. Zhai and A. Barreto, "Stress Recognition Using Non-Invasive Technology," *Proc. 19th Int'l Florida Artificial Intelligence Research Soc. Conf.*, pp. 395-400, 2006.
[14] J. Stroop, "Studies of Interference in Serial Verbal Reactions," *J. Experimental Psychology*, vol. 18, no. 6, pp. 643-662, 1935.
[15] G. Sakr, I. Elhajj, H. Huijer, C. Riley-Doucet, and D. Debnath, "Subject Independent Agitation Detection," *Proc. IEEE/ASME Int'l Conf. Advanced Intelligent Mechatronics*, pp. 200-204, 2008.
[16] G. Sakr, I. Elhajj, and U. Wejinya, "Multi Level SVM for Subject Independent Agitation Detection," *Proc. IEEE/ASME Int'l Conf. Advanced Intelligent Mechatronics*, pp. 538-543, 2009.
[17] R. Dishman, Y. Nakamura, M. Garcia, R. Thompson, A. Dunn, and S. Blair, "Heart Rate Variability, Trait Anxiety, and Perceived Stress among Physically Fit Men and Women," *Int'l J. Psychophysiology*, vol. 37, no. 2, pp. 121-133, 2000.
[18] J. Lee, F. Pearce, A. Hibbs, R. Matthews, C. Morrissette, and R. Walter, "Evaluation of a Capacitively-Coupled, Non-Contact (through Clothing) Electrode or ECG Monitoring and Life Signs Detection for the Objective Force Warfighter," http://ftp.rta.nato.int/public/Pubfulltext/RTO/MP/RTO-MP-HFM-109///MP-HFM-109-25.pdf, 2010.
[19] M. Malik, J. Bigger, A. Camm, R. Kleiger, A. Malliani, A. Moss, and P. Schwartz, "Heart Rate Variability: Standards of Measurement, Physiological Interpretation, and Clinical Use," *European Heart J.*, vol. 17, no. 3, p. 354, 1996.
[20] D. Murray, "What Is Heart Rate Variability; and Is It Blunted by Tumor Necrosis Factor?" *Chest*, vol. 123, no. 3, p. 664, 2003.
[21] A. Kistler, C. Mariauzouls, and K. von Berlepsch, "Fingertip Temperature as an Indicator for Sympathetic Responses," *Int'l J. Psychophysiology*, vol. 29, no. 1, pp. 35-41, 1998.
[22] K. Pearson, "LIII. On Lines and Planes of Closest Fit to Systems of Points in Space," *Philosophical Magazine, Series 6*, vol. 2, no. 11, pp. 559-572, 1901.
[23] V. Kecman, *Learning and Soft Computing: Support Vector Machines, Neural Networks, and Fuzzy Logic Models.* MIT Press, 2001.
[24] N. Aronszajn, *Introduction to the Theory of Hilbert Spaces.* Oklahoma Reasearch [sic] Foundation, 1950.
[25] V. Vapnik, *Statistical Learning Theory.* Wiley, 1998.
[26] C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 121-167, 1998.
[27] C. Spielberger, R. Gorsuch, and L. Edward, *STAI Manual for the State-Trait Anxiety Inventory ("Self-Evaluation Questionnaire").* Consulting Psychologists Press, 1970.

George E. Sakr received the BE degree (with distinction) in electrical and electronics engineering from Lebanese University, Roumieh, in 2005, and the MS degree in networking and telecommunications from the joint program between Lebanese University, the Universite de Versailles St. Quentin, the Ecole Nationale Superieure des Telecommunications (ENST, France), Universite Pierre et Marie Curie (UPMC, France), and the Institut National de la Recherche Scientifique (INRS Telecommunications, Canada), in 2006. He is currently working toward the PhD degree in the Department of Electrical and Computer Engineering at the American University of Beirut. His research interests include artificial intelligence, statistical learning theory, manifolds, human machine interfacing, and medical systems. He is a member of the IEEE.

Imad H. Elhajj received the BE degree (with distinction) in computer and communication engineering from the American University of Beirut (AUB), Lebanon, in 1997, and the MS and the PhD degrees in electrical engineering from Michigan State University, East Lansing, in 1999 and 2002, respectively. He is currently an assistant professor with the Department of Electrical and Computer Engineering, AUB. His research interests include instrumentation, sensor and computer networks, robotics, human machine interfacing, multimedia networking, and medical systems. He is a senior member of the IEEE and the secretary for the IEEE Lebanon Section.

Huda Abou-Saad Huijer received the BSN degree from the American University of Beirut, Lebanon, in 1971, and the master's and PhD degrees from the University of Florida in 1975 and 1977, respectively. She is currently the director of Hariri School of Nursing at the American University of Beirut. She has served in different capacities on a large number of national and international organizations such as the International Association for the Study of Pain, the European Association for Palliative Care, and the European Academy of Nursing Science; of the latter, she was vice-president. In Lebanon, she is currently the president-elect of the Lebanese Society for the Study of Pain, and an active member of the Lebanese Pain Relief and Palliative Care Working Group. Her research interests include pain management and palliative care in children and adults.