Classifying Stress From Heart Rate Variability Using Salivary Biomarkers as Reference

Wei Shiung Liew, Manjeevan Seera, Senior Member, IEEE, Chu Kiong Loo, Senior Member, IEEE, Einly Lim, and Naoyuki Kubota, Member, IEEE

Abstract—An accurate and noninvasive stress assessment from human physiology is a strenuous task. In this paper, a pattern recognition system to learn complex correlates between heart rate variability (HRV) features and salivary stress biomarkers is proposed. Using the Trier social stress test, heart rate and salivary measurements were obtained from volunteers under varying levels of stress induction. Measurements of salivary alpha-amylase and cortisol were used as objective measures of stress, and were correlated with the HRV features using fuzzy ARTMAP (FAM). In improving the predictive ability of the ARTMAPs, techniques, such as genetic algorithms for parameter optimization and voting ensembles, were employed. The ensemble of FAMs can be used for predicting stress responses of salivary alpha-amylase or cortisol using heart rate measurements as the input. Using alpha-amylase as the stress indicator, the ensemble was able to classify stress from heart rate features with 75% accuracy, and 80% accuracy when cortisol was used.

Index Terms—Alpha-amylase, cortisol, fuzzy ARTMAP (FAM), genetic optimization, heart rate variability (HRV), negative correlation (NC), probabilistic voting.

I. INTRODUCTION

The human body responds to stress by producing a variety of corticosteroids and hormones. Prolonged exposure to stress, or chronic stress, causes a number of physical and mental ailments, including hypertension, cardiovascular disease, and depression. Workload stress is one of the major contributors toward productivity loss and medical costs, especially in highly competitive and high-risk work environments [1]. For efficient management of stress in the workplace, there is a need for a reliable method for automated and real-time monitoring of a person’s stress levels in the natural environment.

The stress monitoring method should be ubiquitous and noninvasive, allowing the monitored subjects to perform their daily activities without impedance and inconvenience. Liao et al. [2] proposed a method in which stress was inferred using noninvasive means, using visual behavioral and physiological cues as input to a Bayesian inference engine. However, this method relied on a direct visual surveillance at a static position and angle. For this purpose, electrocardiogram (ECG) sensors may be more useful for a mobile monitoring system [3]. Using a Stroop test to manipulate stress in the subjects, the heart rate variability (HRV) features obtained from the recorded ECG displayed a significant difference between the evoked stress state and the baseline state.

In experiments performed in [4], ECG and salivary measurements were obtained from subjects in parallel, using the Trier social stress test (TSST) for inducing different magnitudes of psychosocial stress. While salivary alpha-amylase is shown to be a good biomarker of psychosocial stress, it is not a practical method for continuous and extended stress assessment. In this paper, the fuzzy ARTMAP (FAM) neural network architecture is proposed for pattern learning and correlating HRV features to amylase and cortisol measurements. In [5], using the same data set, we proposed an FAM-based pattern learning and classification system using genetic algorithms (GAs) for parameter tuning. The methodology proposed here augmented the previous work in several ways, including using GA for ensemble creation, and improved diversification of the FAM population using a different GA method, feature subset selection, and negative correlation (NC).

As the performance of FAM is known to be affected by several factors, such as the presentation order of training data [6], GA was proposed in optimizing the ARTMAP parameters. To further improve the recognition rate, multiple FAMs were established in a classifier ensemble using a probabilistic voting strategy in order to combine individual classifier decisions. The configuration of the classifier ensemble is optimized using a combination of GAs and NC. The framework is designed to achieve consistent optimum classification results from the FAM.

The contribution of this paper is twofold. The first contribution is the optimization of FAM using GA ensembles, which improves the presentation order of the training sequence and ARTMAP parameters. The second contribution is the development of an experiment methodology for acquiring heart rate measurements in parallel to salivary measurements under a variety of stress induction techniques. The goal is to develop a useful data set so that the FAM would be able to correlate heart rate measurements with salivary measurements.
and replace salivary stress measurement. The organization of
this paper is as follows. In Section II, a literature review is first
presented, which is followed by the details of the proposed
system in Section III. The experimental setup is elaborated in
Section IV, while the results of the experiment are discussed in
Section V. Finally, the conclusion is drawn in Section VI.

II. LITERATURE REVIEW

In this section, a literature review on various applications,
based on FAM and GA, is presented. Nater et al. [4] performed a study into human salivary alpha-amylase and
cortisol reactivity in a psychosocial stress paradigm. No sig-
nificant correlations were discovered between alpha-amylase
and cortisol, although both measurements along with heart
rate showed marked increases in reaction to stress stimuli.
Filaire et al. [7] studied stress responses in subjects who
were asked to deliver lectures in front of a crowd. HRV and
salivary alpha-amylase were obtained along with a perceived
stress scale survey. Although no link was found between the
survey and the physiological measures, giving lectures resulted
in correlated changes in alpha-amylase and in the high-
frequency HRV features. The study revealed no gender differ-
ences for HRV and alpha-amylase activity. Rimmel et al. [8]
found that physical fitness has a differential effect on stress-
related neurophysiological systems in response to psychosocial
stress. In terms of heart rate, cortisol, and psychological
responses, professional athletes showed attenuated responses.
Amateur athletes displayed lower heart rate but the same
cortisol responses when compared with untrained subjects.
Almela et al. [9] studied the effects of age on salivary response
to psychosocial stress. No age differences were found in the
salivary responses to stress, although older adults displayed
higher output of salivary alpha-amylase.

The FAM neural network introduced in [10] was capable
of learning complex correlations between multidimensional
input patterns and output class categories. Its fast incremental
learning and ability to retain previous knowledge through
successive training makes it a popular choice for pattern learn-
ing and classification studies. Nevertheless, the performance
of the FAM can be degraded through classifier overfitting
and dependence on training sequence. Extensions of the
FAM model were proposed to address various issues, creating
used a more effective internal representation to improve
robustness of the learning system against noisy data sets. The
biased ARTMAP [12] introduced a mechanism for biasing the
internal knowledge representation whenever a predictive error
occurred. The TPPFAM [13] can reduce category prolifera-
tion using the threshold filtering mechanism and improving
classification prediction using a dynamic Q-max rule.

The performance of the FAM is sensitive to the given
sequence of training data during supervised learning. For
a data set with many patterns, any two randomly selected
training sequences may result in FAMs with a significant
difference in classification accuracy. A commonly utilized
method in addressing this issue is to perform training multiple
times with random sequences and averaging the results.
Another method is to search for optimum training sequences.
Dagher et al. [6] used a max–min clustering method to
determine the training order that would generate the best
performance from the FAM. Other methods for finding the
optimum ordering scheme include GAs [14]–[16] and particle
swarm optimization (PSO) [17]. In all cases, the performance
of the ARTMAPs on average was better than random training.

GAs are multiparameter search heuristics mimicking natural
evolutionary processes to evolve good solutions for a given
problem. In this case, GA is employed to determine the
optimum combination of parameter settings to create and train
an FAM to achieve better classification accuracy. Aside from
optimizing the training order, the GA also sets the values
for ARTMAP parameters, such as baseline vigilance and
learning rate, as outlined in [18]. In [19], given a population
of classifiers, a multiobjective GA was used to select a subset
of the population to assemble into a classifier ensemble. The
proposed ensemble solutions were evaluated on the basis of
diversity and ensemble accuracy. Another similar method used
in [20] studied four different measures for evaluating each
ensemble generated by the GA. Sylvester and Chawla [21]
used a different approach using the GA to weight the contri-
butions of each classifier in the ensemble, instead of assuming
equal contributions from each individual. Kaylani et al. [22]
used GA to directly optimize the topology and architecture of
the ARTMAP neural networks.

The main issue in [5] was the underwhelming improvement
in performance when multiple FAMs were combined in an
ensemble by manually selecting the top-performing classifiers
from the population. However, the issue of ensemble diversity
was not accounted for. In the ensemble theory, diversity
was generally defined as the qualitative difference in output
between each classifier in the ensemble. Ruta and Gabrys [23]
and Kuncheva and Whitaker [24] each performed a survey to
link diversity to ensemble accuracy, but no definitive conclu-
sions were achieved. Nevertheless, an ensemble of classifiers
with uncorrelated errors would often produce better accuracy
than classifiers with correlated errors, due to the implementa-
tion of the voting system.

Increasing the performance of the ensemble of classifiers,
therefore, involved improving the overall diversity of the
population of FAMs. The hierarchical fair-competition paral-
lel GA (HFCPGA) [25] was able to generate more diverse
solutions by distributing the search across multiple subpopu-
lations [26]. In addition, a feature subset selection was also
used to implement specialization by having each classifier trained
using different subsets of features [27]. Having generated a
population of diverse individual FAMs, the next step was to
select classifiers for the ensemble. In [28], an ensemble with
a subset of classifiers may produce better accuracy than using
the entire population. Searching for the best combination of
FAMs for the ensemble can be done using another iteration of
the GA [19]. NC [29] was used to assemble the classifier
ensemble to ensure minimal overlap between the FAMs.

III. PROPOSED SYSTEM OVERVIEW

In this section, an overview of the proposed system is
detailed. Given a pattern classification task with a data set of
labeled examples, the algorithm first generates a population
Fig. 1. Flowchart of FAM optimization using GAs.

of candidates, or chromosomes, each of which can be used to create a singular trained FAM pattern classifier. The hierarchical parallel genetic optimization method is used to search for optimum chromosomes by competitive eliminations and generating new variants from the survivors. Subsequently, another GA is performed to select a number of candidate solutions from the last population to form an ensemble of FAM classifiers, whose fitness is calculated as an aggregate of classification accuracy and interclassifier diversity. The final ensemble of classifiers can then be used as a pattern classifier, utilizing probabilistic voting to combine individual classifier decisions. A flowchart of the process is shown in Fig. 1. The individual modules are described in detail.

A. Fuzzy ARTMAP

The basic FAM architecture consists of the input layer, the output layer, and an intermediate layer. Training the FAM for pattern classification involves supervised learning with a set of labeled exemplars. During the training phase, an exemplar was presented to the input layer, while the output layer receives the category label representing the correct class prediction given that particular exemplar. A mapping process adjusts the weightage of the nodes in the intermediate layer until the input can be matched to the output with a minimal adjustment of the node weights. A user-defined parameter, vigilance, determines the degree of flexibility allowed during the matching process. Setting a low vigilance value allows the FAM to learn generalized, abstract concepts, while a high vigilance increases specialization.

Each discrete class category can be represented as a multi-dimensional hyperbox in the solution space, while a single exemplar can be represented as a point in the multidimensional space. As each exemplar is presented for training, if its position lies at the outside of the hyperbox, minimal adjustments were made to expand the boundaries of the hyperbox to include the exemplar. The rate of expansion is limited by an FAM parameter, vigilance. Zero vigilance allows unlimited expansion, and thus enabling greater generalization by having the hyperbox cover a larger area of the feature space.

The nature of the hyperboxes in the FAM architecture means that the sequence of training exemplars presented to the classifier will influence how the hyperbox will expand. Two different training sequences may result in two distinct FAM classifiers, each with a different classification performance [12]. Thus, for any given training data set, there exist one or more sequences that will yield a classifier with the best classification accuracy. In addition, Granger et al. [18] proposed tuning several FAM parameters to improve the performance of the neural network. The aforementioned baseline vigilance establishes the starting value for the vigilance parameter for governing the classifier’s generalization capability. Other parameters are the choice parameter, the rate of learning, and match tracking parameter. Similar to the sequencing problem above, determining the optimal combination of parameter settings is an optimization problem. As stated in the literature review, several variants
of the basic ARTMAP architecture have been developed to address the above issues, including Gaussian ARTMAP [11], biased ARTMAP [12], and TPPFAM [13].

B. Hierarchical Parallel Genetic Algorithms

GAs are search algorithms using principles of biological evolution to obtain solutions to a given problem [30]. A potential solution to the problem is encoded in the form of a string of numbers, known as a chromosome. A population of chromosomes is maintained over the course of multiple generations, in which a series of competitive eliminations and breeding genetic variants of the surviving chromosomes ensure that the population of solutions improves incrementally over time.

One problem usually encountered, especially in GAs with high turnover in each generation, is that of convergence. With each successive round of elimination and reproduction, the number of chromosomes sharing common genetic traits increases, eventually leading to the population clustering around a relatively small area of the entire spectrum of possible solutions. In the context of ensembles, diversity between classifiers is important for creating an ensemble that can significantly outperform any single classifier. The default GA is, therefore, unsuitable for optimizing a population of FAMs, since there is a high likelihood of premature convergence, leading to homogeneity among the solutions.

The HFCPGAs [25] are conceived as a technique for mitigating the solution convergence problem. The HFCPGA employs multiple subpopulations of chromosomes, each developing and evolving independently of each other. Each subpopulation groups chromosomes of similar fitness, creating a hierarchy in which the best chromosomes occupy the top subpopulation, down to the least fit chromosomes inhabiting the lowest subpopulation. The multitude of smaller communities allows parallel searching in multiple areas of the solution space, as well as promoting global diversity by splitting the genetic convergence across multiple subpopulations. Periodic migration allows high-fitness chromosomes to move to a higher hierarchy, and vice versa, injecting new diversity into each converging subpopulation.

A single chromosome, as shown in Fig. 2, is encoded with all the information required to create an initial FAM classifier and subsequently train it using a sequence of exemplars with the predefined feature subset. The three parts, which constitute a chromosome, are detailed.

1) Training Sequence: Given a data set with \( N \) exemplars, this string is a random sequence of numbers from 1 to \( N \) representing the order in which the exemplars will be presented to the FAM for training.

2) ARTMAP Structural Parameters: Choice parameter \( \alpha \), learning rate \( \beta \), match tracking parameter \( \epsilon \), and baseline vigilance \( \bar{\rho} \). The parameter settings are randomized according to the range defined in [18]: \( \alpha > 0, \beta \in [0, 1], 0 < \epsilon \ll 1, \) and \( \bar{\rho} \in [0, 1] \).

3) Feature Subset Selection: Given that each exemplar consists of \( M \) attributes or features, this \( M \)-length binary string decides which feature to keep and which to discard. Subsequently, all exemplars in the data set will be formatted following this arrangement. A user-defined parameter can be implemented to ensure a minimum or a maximum limit to the number of selected features in a single chromosome.

The process is outlined as follows.

1) A population of a set number of random chromosomes is generated. The given data set of \( N \) exemplars are indexed sequentially from 1 to \( N \). This is considered as the original sequence of training data.

2) Each chromosome is used to generate a single FAM classifier. First, the structural parameters are used for creating the initial FAM neural network. The exemplars in the original training data set are rearranged following the order of the defined sequence, and reformatted to discard the deselected attributes.

3) The rearranged training data set is then used for training and testing the FAM, using tenfold cross validation.

4) Chromosomes are grouped evenly into subpopulations in order of fitness. A predefined number of the least fit chromosomes are discarded.

5) For each discarded chromosome, a new offspring chromosome is generated using the following method, where an inverse roulette-wheel method selects a subpopulation from which two parent chromosomes are chosen at random to generate a new offspring chromosome. Subpopulations with low-fitness chromosomes will be chosen more often than high-fitness subpopulations as a method for encouraging exploration into newer areas of the solution space rather than converging on the local maxima. Each chromosome is created by genetic crossover and mutation, as shown in Fig. 3.

6) In each generation, any chromosome, which scores a higher fitness than its colleagues in the same subpopulation, is moved up into the next subpopulation. Likewise, a chromosome with poor performance is moved into a lower subpopulation. Only a limited number of migrations are allowed per generation, reserved only for chromosomes with the highest difference between its fitness and its subpopulation’s average fitness.

7) The generation counter is incremented, and steps 2–6 are repeated until a stopping criterion is fulfilled either by reaching a maximum generation limit, or until a number of consecutive generations elapse with no improvement in the population.

Two stopping criteria were defined to terminate the optimization. One of the conditions is when a maximum number of generations have elapsed. At the end of each generation, the mean fitness and the best fitness of the population of FAMs were recorded. The convergence criterion terminates the optimization before reaching the maximum number of generations if the mean and best fitness did not improve over several consecutive generations.
Fig. 3. Offspring chromosome created by inheriting traits from parent chromosomes. Different parts of the chromosome used different methods for combining genes from the parents.

A new offspring chromosome is created by genetic crossover of two parent chromosomes, which is followed by genetic mutation. Crossover and mutation operations were performed on each of the three subsections of the chromosomes independent of other sections. The completed subsections are then reconstituted into a single new chromosome. Fig. 3 shows the methods used for creating the offspring chromosome. The details are as follows.

1) For the training sequence, genes common to both parents are passed down to the offspring. For the remaining uncommon genes, each parent has equal probability of passing down their gene to the offspring. A separate array of numbers keeps track of which genes have been transferred to the offspring in order to prevent duplicates and missing genes. Mutation of this subsection involved swapping the position of any two randomly selected genes.

2) ARTMAP structural parameters $\alpha$, $\beta$, $\epsilon$, and $\bar{\rho}$ for the offspring chromosome is calculated as the mean of the parents’ parameters.

3) For the feature subset selection, logic AND is applied. Genes that have the same value in both parents are passed down to the offspring. Otherwise, the offspring gene has 50% chance to be either 0 or 1. Mutation in this subsection involved flipping the gene between the two values.

4) The three portions are then recombined into a single string, resulting in the offspring chromosome.

C. Ensemble Creation

In Section III-B, a population of trained FAM classifiers was created. A classifier ensemble can be constructed by selecting and combining individual FAMs, ideally with a degree of interclassifier diversity and uncorrelated predictive errors. Fig. 4 shows the entire process of generating optimized ensembles from the given population of FAMs. The NC method [29], [31] is implemented to encourage the selection of classifiers with low correlation with each other to maintain diversity and generalization.

Given a classification task for a set of data $D$ with $N$ exemplars

$$D = \{(A_i, b_i) | i = 1, \ldots, N\}. \quad (1)$$

The data set consists of an array of feature vectors $A_i$ and a category label $b_i$. The $j$th neural network and its $k$th output can be denoted as

$$f_j = [0, 1]^C \quad (2)$$

$$f_j^k = [0, 1] \quad (3)$$

where

$$f_j = (f_j^1, \ldots, f_j^C). \quad (4)$$

The output for an ensemble with $J$ classifiers is computed as

$$f^k(x_n) = \frac{1}{J} \sum_{j=1}^{J} f_j^k(x_n). \quad (5)$$
In the context of NC, the generalization error of an ensemble with \( J \) classifiers is
\[
\hat{E}_J = N \sum_{n=1}^{N} \sum_{k=1}^{C} \left[ (f_k^j(x_n) - y_n^k)^2 \right] - \lambda \sum_{r=1}^{J} \left[ (\hat{f}_k^r(x_n) - f_k^r(x_n))^2 \right].
\]  
(6)

The above generalization error equation can be reduced to
\[
\hat{E}_J = U - \lambda V.
\]  
(7)

The term \( U \) is the predictive error of the resultant classifier ensemble, and \( V \) is the sum of interclassifier diversity. \( \lambda \) in the equation represents a user-controlled parameter for the diversity penalty terms for classifiers, which are similar to one another. Setting \( \lambda = 0 \) reduces the \( \hat{E}_J \) to only the predictive error of the classifiers.

The GA for optimizing classifier selection used a binary string for selecting and deselecting classifiers in the population, and a random fractional number ranging from 0 to 2 for the diversity parameter \( \lambda \). Given a selection of classifiers, the ensemble is initiated with a single classifier. The NC index is calculated, \( \hat{E}_1 \). Subsequently, the next selected classifier was added into the ensemble, and the resultant ensemble’s NC index is computed, \( \hat{E}_2 \). If the newly added classifier does not contribute in terms of ensemble accuracy or diversity (i.e., if \( \hat{E}_2 \geq \hat{E}_1 \)), then the classifier is removed from the ensemble. Using the GA, optimum classifier ensembles are derived for maximum interclassifier diversity and minimum ensemble predictive error.

**D. Probabilistic Decision Fusion**

The ensemble decision from the multitude of selected classifiers is decided using a form of probabilistic voting \([32], [33]\). In general, the probability of classifier \( j \) correctly classifying a given input \( A_i \) is \( p_j \). In case of misclassification \( (e_j = 1 - p_j) \), it is assumed that the \( C - 1 \) residual class categories have equal probability of being selected instead of the correct class \( b_i \)
\[
P(f_j(A_i) = b_i) = p_j
\]  
(8)
\[
P(f_j(A_i) = C_m) = \frac{1 - p_j}{C - 1} = e_j
\]  
(9)

where
\[
i = 1, 2, \ldots, M
\]  
\[
j = 1, 2, \ldots, J
\]  
and
\[
C_m \neq b_i.
\]

Assuming each classifier decides independently of other classifiers, the decision function assumed Bayes’ rule selection to select the class category \( C_m \) with the largest \( a \) posteriori probability in order to minimize the error rate of
the classifier combination. The decision function, therefore, can be summarized as

\[
D_m(A_i) = \ln P(f_j(A_i) = C_m) + \sum_{i=1}^{J} \ln \left( \frac{p_j}{e_j} \right) \delta_{im}(A_i)
\]

\[
= \ln P(f_j(A_i) = C_m) + \sum_{i=1}^{J} \ln \left( \frac{(C-1)p_j}{1-p_j} \right) \delta_{im}(A_i)
\]

where

\[
\delta_{im}(A_i) = \begin{cases} 1, & \text{if } f_j(A_i) = C_m \\ 0, & \text{if } f_j(A_i) \neq C_m \end{cases}
\]

Assuming an ensemble of \( J \) classifiers \{\( f_1, \ldots, f_i, \ldots, f_J \)\} for a classification task in which a given input feature vector \( A_i \) is classified into one of \( C \) discrete class categories. The class \( C_m \) that maximizes the decision function \( D_m(A_i) \) is then selected as the ensemble consensus. Each classifier \( f_j \) is weighted according to its predictive accuracy \( p_j \), and each class category \( C_m \) has a constant representing its \( a \ priori \) probability.

The use of this method allows an uneven voting scheme, whereby each member of the ensemble is given priority according to its predictive accuracy. The final decision of the ensemble is therefore more skewed toward reliable classifiers. In the case when the reliable classifiers make a classification error, the correct decision may still be selected with enough support from lesser classifiers.

IV. EXPERIMENTAL SETUP

The experiment is performed for testing the viability of a neural network that can mimic human salivary responses for the alpha-amylase and cortisol. Using the Trier social stress test for stress induction, ECG and salivary measurements were obtained from a number of subjects under different stress conditions. HRV features were extracted from the ECG recordings. Alpha-amylase and cortisol measurements were derived from the saliva samples, and were used to approximate the level of stress. FAM was then used to establish pattern correlations between the extracted HRV features to the approximated stress levels. When completed, the FAMs can be used to approximate salivary stress responses given any ECG signal recordings.

A. Trier Social Stress Test

The Trier social stress test [34] is a protocol for psychological stress induction in a laboratory setting. A single session typically consisted of an anticipation period followed by a test period in which the subject has to deliver a free speech and perform mental arithmetic before an audience. The protocol has been found to induce significant changes in the concentration of adrenocorticotropin hormone, cortisol, growth hormone, prolactin, and HRV.

All subjects were volunteers recruited among the student population of a university. Volunteers were first vetted for factors that might introduce a bias in the experiment, such as smoking and drinking habits, preexisting health conditions, and medication intake. Due to the strict vetting process as well as time and resource constraints, only 22 subjects qualified for the data collection task. There were 17 male subjects and 5 females, with a mean age of 21 years. Subjects were informed that they would be performing two sessions: 1) one interview session and 2) one reading session, along with mental arithmetic tasks to simulate stress. The choice to perform control session or stress session first was determined at random. Each session was scheduled on different days, no more than one week in between sessions. Subjects were informed to fast for at least 12 h before participating in each session. They were also informed regarding the physiological measurements and have signed consent forms to use their measurements in a research capacity.

1) Experiment Protocol: Three bipolar silver-chloride electrode pads were used in a modified lead-2 configuration. An electrode was attached to the inside of both wrists, and a third was grounded to one ankle. The signal acquisition device is g.Mobilab developed by g.Tec. The device was attached near the participant’s seat where it wirelessly transmitted the continuous signals from the electrodes to a nearby computer at a 256-Hz sampling rate.

Salivary samples were obtained using swabs from salivary immunoassay kits, following the standard guidelines outlined in the manual provided by the manufacturer, Salimetric Inc. The experiment was performed in an air-conditioned room, between 2 P.M. and 6 P.M. to eliminate diurnal variations. The subject was seated comfortably, facing a panel of two interviewers five feet away. One interviewer manipulated the computer interfacing with the g.Mobilab data acquisition device and was responsible for sorting the digital recordings for each session. Each session consisted of five parts, as shown in Fig. 5.

1) Pretest: During this period, the electrodes were attached to the subject, and the digital signal acquired by the computer was inspected before proceeding with the rest of the session. During this time, the subject was briefed by a panel member regarding the methodology for the current session.

2) Anticipation: For the stress session, the subject was given 3 min to prepare for a mock job interview. In the case of control session, the subject skimmed through an article of text selected for the experiment. For the whole 3 min, the subject’s heart rate was recorded (ECG1). At the end of the duration, a saliva sample was acquired (AC1) by instructing the subject to chew on the swab for \( \sim 10 \) s.

3) First Task: For the stress session, the subject presented their qualifications in a mock job interview. The panel members performed the roles of interviewers by interrogating the subject with standard job interview questions. For the control session, the subject was only required to read aloud the given article with no feedback from the panel. ECG2 was recorded for the duration, and AC2 was acquired at the end.

4) Second Task: For the stress session, the subject was given an arbitrarily high number and was instructed to
serially subtract 7 from the given number as quickly and as accurately as possible. Whenever a mistake was made or if the subject slowed down, they were instructed to start again with a new number. In the case of the control session, the subject instead serially added 10 from 0. ECG3 and AC3 were obtained.

5) Posttest: During this time, the subject was given a visual analog scale questionnaire to assess the relative discomfort levels of the current session. ECG4 was recorded for the first 5 min, while saliva sampling was performed at 10 min intervals for AC4, AC5, and AC6.

2) Data Preprocessing: A total of 176 ECG recordings and 264 salivary samples were obtained from 22 subjects during the course of the data collection experiment. ECG recordings were stored in g.Tec’s proprietary digital format on a computer hard drive, while the salivary samples were stored in a laboratory refrigerator. Six salivary samples were rejected due to insufficient volume. All saliva samples were centrifuged at 3000 r/min for 15 min. Optical density reading was performed using a microtitre plate spectrophotometer by Biotek, along with a GEN5 Microplate Data Collection and Analysis software. Measurement was performed twice using duplicate samples, and the results were averaged.

Each ECG segment was normalized to eliminate baseline and individual variances. Feature extraction was performed on each ECG segment in order to reduce dimensionality, using the Augsburg Biosignal Toolbox [35] in addition to the principal dynamic modes method [36] for decomposing the ECG vector. Each ECG segment was thus represented by 106 features consisting of time and frequency domain as well as nonlinear dynamic components. The principal dynamic modes features, in particular, displayed certain characteristics suitable for nonlinear stress analyses, as demonstrated in [37].

Analysis of variance (ANOVA) derived the statistical difference of salivary measurements between control and stress sessions ($F = 18.08$ and $p = 3.04e-5$). A three-category classification task was proposed, using the alpha-amylase and cortisol measurements as reference to stress levels. Based on the statistical analyses, the high cutoff for the low stress category used the mean measurement minus one standard deviation, while the low cutoff for high stress was the mean measurement plus one standard deviation. The normalized ranges for the three-class label for the alpha-amylase and cortisol were derived in Table I. Normalization was performed within-subject for measurements obtained from both stress and control sessions to negate baseline and individual variances.

Fig. 6 shows the averaged and normalized measurements of the alpha-amylase and cortisol for control and stress sessions. Alpha-amylase showed immediate reaction to the stress stimuli as well as a quick return to baseline after the first 10 min of rest. However, we could not account for the huge fluctuations of alpha-amylase measurements during the control sessions. Cortisol measurements during the control sessions are relatively stable, with a slight increase due to the stress of performing even relatively simple tasks. Measurements during the stress sessions showed a marked increase even during through the first rest period, and then stabilized and slowly decreased for the remaining rest periods. We hypothesize that salivary cortisol may have a slight lag in showing responses to stress stimuli.

The first data set used the alpha-amylase measurements to group ECG feature vectors into three distinct classes representing its objective stress levels. The second data set used the cortisol measurements instead to label its ECG feature vectors.

<table>
<thead>
<tr>
<th>TABLE I</th>
</tr>
</thead>
<tbody>
<tr>
<td>NORMALIZED RANGES OF ALPHA-AMYLASE AND CORTISOL MEASUREMENTS IN EACH CLASS CATEGORY, DERIVED USING ANOVA</td>
</tr>
<tr>
<td><strong>Salivary measurement</strong></td>
</tr>
<tr>
<td>Alpha-amylase</td>
</tr>
<tr>
<td>Cortisol</td>
</tr>
</tbody>
</table>
Fig. 6. Averaged and normalized alpha-amylase and cortisol measurements for each section for a single TSST session.

TABLE II
CORRELATING ECG AND SALIVARY MEASUREMENTS WITH VARYING LEVELS OF LATENCY

<table>
<thead>
<tr>
<th>Data set</th>
<th>Latency=0</th>
<th>Latency=1</th>
<th>Latency=2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anticipation</td>
<td>ECG1 + AC1</td>
<td>ECG1 + AC2</td>
<td>ECG1 + AC3</td>
</tr>
<tr>
<td>First task</td>
<td>ECG2 + AC2</td>
<td>ECG2 + AC3</td>
<td>ECG2 + AC4</td>
</tr>
<tr>
<td>Second task</td>
<td>ECG3 + AC3</td>
<td>ECG3 + AC4</td>
<td>ECG3 + AC5</td>
</tr>
<tr>
<td>Post-test</td>
<td>ECG4 + AC4</td>
<td>ECG4 + AC5</td>
<td>ECG4 + AC6</td>
</tr>
</tbody>
</table>

Furthermore, data sets can be derived under the assumption that there exists a significant delay or latency between the HRV measurements and the corresponding salivary response. The initial data sets assume zero latency by correlating AC1 salivary measurements to ECG1, AC2 to ECG2, and so on, as shown in Fig. 5. Latency was introduced by correlating the HRV measurements to later salivary measurements, as shown in Table II, giving another four data sets. All the data sets used in the experiment are given in Table III. Correlation coefficients were calculated between each of the extracted ECG feature to the salivary measurements, and the average p-values of all features were reported in the table. The first two data sets have zero latency between ECG and salivary measurements, while latency was introduced in the remaining four data sets. Discrepancy in number of instances between data sets was due to insufficient salivary samples.

B. Pattern Learning and Classification

Having generated six data sets in the previous experiment, each data set was used as input to the proposed system to develop an ensemble of FAMs that would receive ECG-derived HRV features and output the approximate salivary stress response. The following steps were performed.

1) A population of randomized chromosomes was created. Each chromosome represented a single configuration for initializing an FAM and trains it using the rearranged data set.

2) Each chromosome was evaluated.
   a) A single FAM was created using the information encoded in the chromosome. ARTMAP parameters were used for initializing a blank FAM, which was then trained and tested using the given data set that was rearranged according to sequence, and filtered according to the feature subset selection.
   b) FAM training and testing was performed using tenfold cross validation. The overall fitness of the chromosome takes the average classification accuracy of the tests.

3) Genetic operations were performed using the steps outlined according to the HFCPGAs [25]. Chromosomes were separated into hierarchies according to fitness, where genetic selection and reproduction were executed in parallel over multiple generations. Parameter settings for the GA are outlined in Table IV.

4) The optimization task was terminated when one of two stopping criteria was achieved, either after a maximum number of generations have elapsed or until several consecutive generations with no improvement in average and a maximum fitness of the population of chromosomes. The second stopping criteria assumed that convergence was achieved and no better chromosomes could be discovered if no further progress was detected in the past few generations.

5) Another GA step was performed to search for the combination of FAMs with the best combined...
ensemble accuracy. Classification output from individual ARTMAPs was integrated using a probabilistic voting method, and assembled using the NC to filter redundant classifiers.

6) The GA was iterated until one of two stopping criteria was fulfilled, either after a maximum number of generations have elapsed or until several consecutive generations with no improvement in average and a maximum fitness of the population of chromosomes.

7) The ensemble with the highest classification accuracy and the smallest number of classifiers were selected.

8) The experiment was repeated ten times, and the results were averaged.

V. RESULTS AND DISCUSSION

The results of the classification experiment are presented in Table V. The average and best classification accuracy of individual FAMs were obtained from the final population after HFCPGA was concluded. After the second GA step, the best ensemble of FAMs was selected for having the highest overall accuracy. The total time taken from start to finish was given in seconds. Ensemble accuracy, sensitivity, and specificity were computed according to Sokolova and Lapalme’s methods for multiclass classification. Results were obtained as an average over ten trials.

In Table III, it is shown that the Cortisol + 0 and Cortisol + 1 data sets have the best feature-to-label correlations in terms of p-values. However, Table V showed that having a good feature-to-label correlation does not equate into a better classifier performance. In terms of individual FAM accuracy, Cortisol + 0 was the best overall data set by having a population of FAMs with a mean accuracy of 65.15%. For Amylase data sets, Amylase + 1 was marginally better than the rest, having a mean accuracy of 59.48% compared with 56.17% and 56.56%. All of the Cortisol data sets scored higher than the Amylase data sets in terms of ensemble accuracy. Again, Cortisol + 0 was the clear winner with both the highest accuracy and the smallest ensemble size. For the Amylase data sets, however, no clear winner presents itself.

The sensitivity and specificity reported here are statistical measures of the performance of the ensemble classification system, averaged across the three classes. Sensitivity is known as the rate of true positives, while specificity is also called the rate of true negatives. In Table V, while Cortisol + 0 was shown to have the highest classification accuracy, Cortisol + 2 has better sensitivity and specificity scores. This may be a case of accuracy paradox, when a predictive model’s accuracy does not represent the true predictive ability of the model, especially if the data set used for training and testing has a skewed class balance. Therefore, statistically, Cortisol + 0 is outperformed by Cortisol + 1 and Cortisol + 2 if accuracy was disregarded in favor of sensitivity and specificity. This appears to be corroborated by the graph of cortisol measurements in Fig. 6, as it shows a slight lag in cortisol responses to stress stimuli.

All cortisol-trained FAM ensembles indicate better performance than amylase-based ensembles in terms of accuracy, sensitivity, and specificity. This may be explained as the Amylase data sets having more correlated misclassifications. For the Cortisol data sets, misclassifications occur as singular anomalies, which can be corrected easily using ensemble voting to select the right outcome.

The limitations of the experiment prevented a more detailed analysis on the behavior of lagged data sets. The Cortisol + 0 data set, in addition to having many HRV features correlated with salivary stress measurements, has the highest accuracy among individual FAMs and the ensemble with the best classification accuracy and the smallest ensemble size.
compared with the other Cortisol data sets. This could mean that Cortisol + 0 represented the data set with the most accurate HRV to salivary measurement mapping, making it an easy task for the ensemble to achieve high accuracy with a relatively few FAMs. In the case of Amylase data sets, it proved difficult in choosing the best data set as the resultant ensembles achieved similar results.

The progression of the optimization process is segmented in Table VI. The second column shows the classification accuracy of the initial population of randomly generated FAMs before optimization was performed. The first stage GA was then employed to optimize the performance of individual FAMs. The third column shows the overall classification accuracy of the population of FAMs after optimization. This population of FAMs was then used by the second stage GA to form classifier ensembles. The initial population was constructed by randomly choosing individual FAMs for each ensemble, and their overall classification accuracy is presented in the fourth column. The last column shows the performance of the ensembles after optimization.

VI. CONCLUSION

In this paper, a method for correlating the human body’s salivary reaction to stress using HRV has been presented. Using a series of stress-induction tests, a data set of salivary and ECG measurements was obtained. An FAM classifier was used for learning and mapping correlations between ECG features and salivary measurements of alpha-amylase and cortisol. The goal was to develop a classifier that can accurately predict the salivary response given a set of ECG measurements by learning from the obtained data set. GAs were employed for parameter optimization and ensemble design to improve the recognition rate. The final classifier system was able to classify salivary cortisol given ECG features with 80% accuracy, compared with 75% accuracy for salivary alpha-amylase.

Processing salivary samples to derive the alpha-amylase and the cortisol as objective indicators of stress is a long and laborious process. The proposed methodology in this paper uses an ensemble of FAMs to learn and approximate the correlations between heart rate responses and salivary responses to stress. The system is then able to predict the appropriate salivary responses using only heart rate measurements as input, far quicker than having to acquire and process each salivary sample. In contrast, heart rate measurements can be obtained in real time and processed within seconds. This system can be implemented as a practical continuous monitoring system to assess stress in patients. While the system has obtained only 80% accuracy in correctly classifying three distinct stress groups, the benefits of near real time and convenient measurement may be considered as a factor to compensate for the reduced accuracy.

In the future work, focus will be placed on real-life medical applications, such as on ECG signals and auscultatory blood pressure signals using our proposed approach. In addition, the framework can be updated to use other optimization methods, such as PSO or Firefly, to design a more diverse classifier ensemble system.

REFERENCES


Wei Shuang Liew received the bachelor’s degree in electronics engineering from Multimedia University–Melaka, Cyberjaya, Malaysia. He is currently pursuing the master’s degree in biomedical engineering with the University of Malaya, Kuala Lumpur, Malaysia.

His current research interests include physiological signal processing, soft biometrics, and machine learning for automated recognition of stress and affect.

Manjeevan Sorea (M’11–SM’15) received the B.Eng. (Hons.) degree in electronics and electrical engineering from the University of Sunderland, Sunderland, U.K., in 2007, and the Ph.D. degree in computational intelligence from Universiti Sains Malaysia, Penang, Malaysia.

He is currently a Senior Lecturer with the Swinburne University of Technology–Sarawak Campus, Kuching, Malaysia. His current research interests include soft computing, fault detection, diagnosis, and human–robot interaction.

Chu Kiong Loo (SM’14) received the B.Eng. (Hons.) degree in mechanical engineering from the University of Malaya, Kuala Lumpur, Malaysia, and the Ph.D. degree in artificial intelligence and robotics from the Universiti Sains Malaysia, Penang, Malaysia.

He is currently a Professor with the University of Malaya. His current research interests include soft-computing, human–robot interaction, Internet of Things, and deep learning.

Einyly Lim received the B.E. (Hons.) and M.EngSc. degrees from the University of Malaya, Kuala Lumpur, Malaysia, in 2003 and 2006, respectively, and the Ph.D. degree from the Graduate School of Biomedical Engineering, University of New South Wales, Sydney, NSW, Australia, in 2009.

She is currently a Senior Lecturer with the University of Malaya.

Naoyuki Kubota (S’95–A’97–M’01) received the B.Sc. degree from Osaka Kyoiku University, Kashiwara, Japan, in 1992, the M.Eng. degree from Hokkaido University, Hokkaido, Japan, in 1994, and the D.E. degree from Nagoya University, Nagoya, Japan, in 1997.

He is currently a Professor with Tokyo Metropolitan University, Tokyo, Japan.