IEEE Journal of Translational Engineering in Health and Medicine

Received 29 August 2016; revised 19 May 2017 and 28 July 2017; accepted 30 July 2017. Date of publication 11 September 2017; date of current version 29 September 2017.

Digital Object Identifier 10.1109/JTEHM.2017.2737992

Toward a Personalized Real-Time Diagnosis in Neonatal Seizure Detection

ANDRIY TEMKO[1], (Senior Member, IEEE), ACHINTYA KR. SARKAR[2], GERALDINE B. BOYLAN[3], SEAN MATHIESON[4], WILLIAM P. MARNANE[1], (Member, IEEE), AND GORDON LIGHTBODY[1], (Member, IEEE)

[1]Department of Electrical and Electronic Engineering and Irish Centre for Fetal and Neonatal Translational Research, University College Cork, T12 P2FY Cork, Ireland
[2]Department of Electronic Systems, Aalborg University, 9220 Aalborg, Denmark
[3]Department of Paediatrics and Child Health and INFANT Center, University College Cork, T12 P2FY Cork, Ireland
[4]Academic Research Department of Neonatology, Institute for Women's Health, University College London, London WC1E 6AU, U.K.

CORRESPONDING AUTHOR: A. TEMKO (atemko@ucc.ie)

This work was supported by the Wellcome Trust Strategic Translational Award under Grant 098983/Z/12 and in part by the Seed Award in Science under Grant 200704/Z/16 and in part by the Science Foundation Ireland Principal Investigator Award under Grant 10/IN.1/B3036. The work of A. Temko was supported by the Science Foundation Ireland through the Research Centres Award under Grant 12/RC/2272.

ABSTRACT The problem of creating a personalized seizure detection algorithm for newborns is tackled in this paper. A probabilistic framework for semi-supervised adaptation of a generic patient-independent neonatal seizure detector is proposed. A system that is based on a combination of patient-adaptive (generative) and patient-independent (discriminative) classifiers is designed and evaluated on a large database of unedited continuous multichannel neonatal EEG recordings of over 800 h in duration. It is shown that an improvement in the detection of neonatal seizures over the course of long EEG recordings is achievable with on-the-fly incorporation of patient-specific EEG characteristics. In the clinical setting, the employment of the developed system will maintain a seizure detection rate at 70% while halving the number of false detections per hour, from 0.4 to 0.2 FD/h. This is the first study to propose the use of online adaptation without clinical labels, to build a personalized diagnostic system for the detection of neonatal seizures.

INDEX TERMS Neonatal, seizure, detection, online adaptation.

I. INTRODUCTION

Individual healthcare decisions [1] empowered by technological solutions such as automatic diagnostic systems have been shown to be more accurate than more generic systems in many areas of biomedical signal processing. These systems are built using the data of a targeted user/patient which eliminates the inter-subject variability of the training data and allows the system to focus on learning intra-subject characteristics, thus simplifying the estimation and recognition problem. Well-known examples include subject-specific brain computer interfaces [2], patient-specific epilepsy detection systems [3] and patient-specific diagnostic consultation However, in the development of an EEG-based seizure detector for the newborn [5], the EEG data of the baby cannot be obtained before the baby is born. The successful system must be able generalize over the pre-recorded and pre-annotated data from other babies to be able to detect seizures from the data of the new baby and alarm the clinical personnel. There are also significant clinical pressures for the availability of useful information from EEG monitoring, within hours of birth.

A number of research groups [5]–[13] have previously developed neonatal seizure detection algorithms (SDA) in an attempt to assist healthcare professionals with objective decision support. A typical SDA comprises of the following main stages: i) The signal representation stage (feature-level) – where relevant features are robustly extracted from the pre-processed EEG signal. ii) The classification stage (classifier level) – where the extracted feature or feature vectors are assigned to the seizure or non-seizure class using a set of rules and thresholds which are either automatically derived from the data (classifier) [5], [8], [13] or manually adjusted following or mimicking the reasoning of expert neurologists [6], [7], [9], [12]. iii) The post-processing stage (decision level) – this involves both temporal smoothing to reduce noise and possibly other transformations that may offer some support in the decision making process to a clinician.

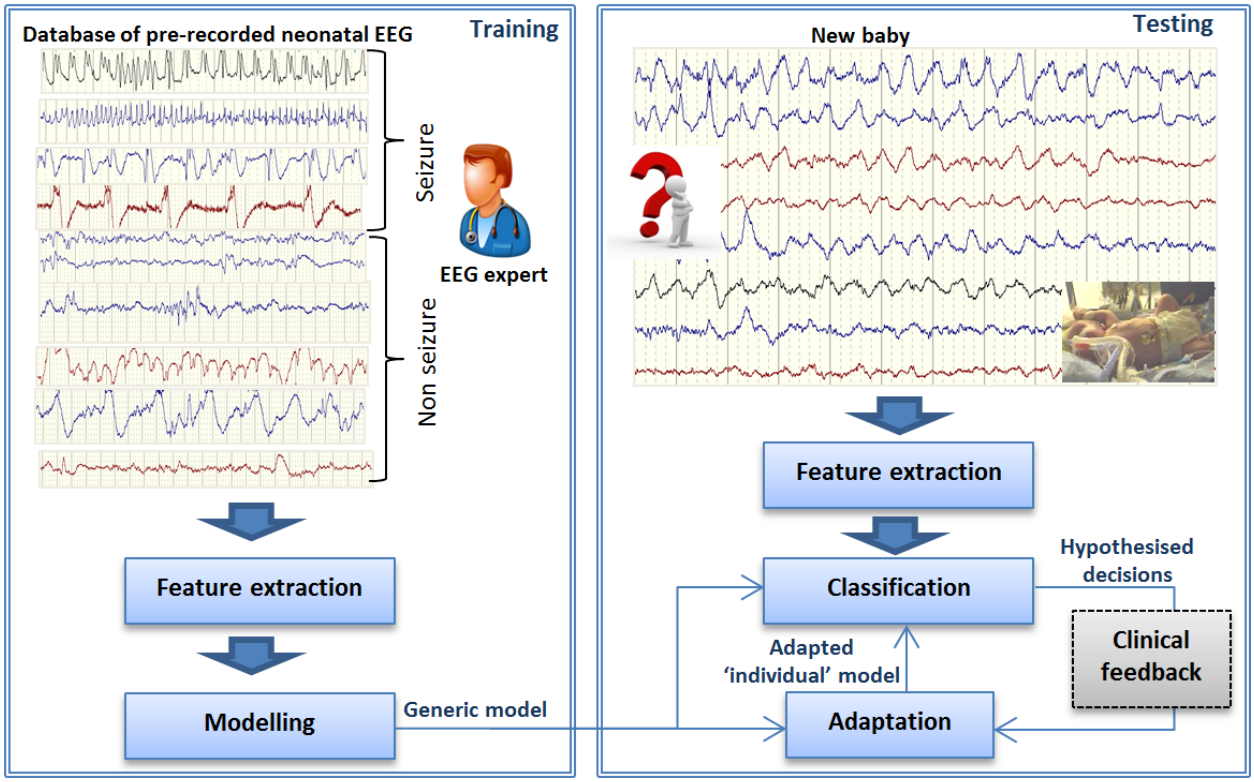

FIGURE 1. From generic to personalized neonatal seizure detector. The adaptation of the generic model can be performed on the fly. The optional clinical feedback in the testing stage can be used to purify the hypothesised decisions.

A notable improvement has recently been achieved in neonatal seizure detection with the development of a Support Vector Machine (SVM) based neonatal seizure detector [14], [15] which has completed a pre-market European multi-centre clinical investigation[1] to support its regulatory approval and clinical adoption [16]. This system utilises an SVM classifier trained on a high dimensional set of extracted features that carry temporal, frequency, structural and energy information about the neonatal EEG. The analysis of algorithmic performance in [17] and [18] revealed that the performance of the detector is significantly correlated with seizure duration, amplitude, rhythmicity and the number of EEG channels involved in the seizure during peak seizure activity.

The combination of various (even many) classifiers has been widely researched both theoretically in the literature [19] and practically through public competitions such as Kaggle [20]. The underlying principle here is that performance improvement may be obtained from the diversity of classification methods; popular classifier combinations include blending, bagging, boosting, stacking, etc. Such classifier ensembles tend to yield good results when there is a significant diversity among the models. This diversity can come from inherent algorithmic randomness (like random decision trees) or from a deliberate difference in the optimisation functions used in training (such as the difference between the discriminative SVM and the generative GMM) or the usage of different training datasets. For example, consider a situation in which there are two seizure classifiers; to make this conceptually easy, consider that one classifier utilises the EEG and the other uses video. If the EEG based classifier just misses a seizure, whilst the other classifier confidently identifies a seizure based on some video cues, it would seem sensible that given its confidence, the video based classifier becomes the expert at this particular moment and can therefore guide the EEG based classifier to better performance over this data. Each classifier is looking at the problem in a different way and may contribute complementary expertise.

In neonatal intensive care units (NICU), neonates that are suspected of developing neurological complications can be continuously monitored using EEG for several days; indeed, pre-term infants can often be monitored over a period of a few weeks. Seizure characteristics can vary between neonates and thus long EEG recordings can be exploited, as shown in Fig. 1, to derive individualized models from generic patient-independent systems. Such a subject adaptive system must capture the specifics of the monitored neonate on-the-fly in order to improve its performance; however, data annotations are required to drive this optimisation. One way to achieve this is to utilize clinical feedback for the small number of suspected events (alarms) and re-build or adapt

[1]https://clinicaltrials.gov/ct2/show/NCT02160171, https://clinicaltrials.gov/ct2/show/NCT02431780

IEEE Journal of Translational Engineering in Health and Medicine

the model to each specific newborn, incorporating the new annotated information. EEG experts are generally not available during unsociable hours and NICU staff typically lack EEG training and many feel unsupported in the interpretation of neonatal EEG [21]. Since the purpose of building the automated SDA is to provide continuous objective brain monitoring that will typically produce alerts when the clinical expertise is not available – this solution may not be practical. An alternative solution is to allow the detector to learn from its own decisions, balancing the gain obtained from new patient-specific information with the uncertainty of its automated hypothesised labels.

A patient-adaptive neonatal SDA is proposed here which combines generic and personalized seizure detectors. Previous work has demonstrated that different classifiers (even when trained on the same data) can provide useful complementary performance to each other for the neonatal patient independent seizure detection task. The novelty of this current study is that a patient adaptive classifier is proposed that will adapt and improve its performance to a specific neonate over the first hours after birth, without the need for clinical labels. A pre-trained patient independent SVM-based system that was developed in [5] and [14] is used to automatically provide labels for data from a new unseen baby which can then be utilised to adapt a Gaussian Mixture Model (GMM) [22] based detector to achieve improved personalised performance, in real time. To the best of our knowledge, this work provides the first application of adaptive personalised neonatal seizure detection without the need for clinical input – this provides state of the art performance for neonatal seizure detection.

II. MATERIALS AND METHODS

A. DATASET

The database is composed of EEG recordings from 18 full-term newborns recruited in the Neonatal Intensive Care Unit (NICU) of Cork University Maternity Hospital (CUMH), Cork, Ireland. The CareFusion NicOne video EEG monitor was used to record multi-channel EEG at 256Hz using the modified 10-20 system of electrode placement with the following 8 EEG bipolar channels F4-C4, C4-O2, F3-C3, C3-O1, T4-C4, C4-Cz, Cz-C3 and C3-T3. All electrographic seizures were annotated independently by an experienced neonatal neurophysiologist (GB) using simultaneous video EEG. The combined length of the EEG recordings totalled 816.7 hours with per patient mean/median length of 45.4/48.5 hours and contained 1389 electrographic seizures. The dataset contains a wide variety of seizure types, including both electrographic-only and electro-clinical seizures of focal, multi-focal and generalized types. The continuous EEG recordings were not manually edited to remove the large variety of artifacts and poorly conditioned signals that are commonly encountered in long EEG recordings in the real-world NICU environment. An additional dataset of 55 non-seizure babies (1 hour per baby) is used to augment the representation of the EEG background activity. This small dataset is used only for training. The described dataset is used to evaluate the developed algorithms retrospectively.

For performance evaluation, the leave-one-out (LOO) procedure was followed where all but one patients' data are used for training and the remaining patient's data are used for testing. The procedure is repeated until each of the 18 patients has been a test subject and the mean results are reported. LOO is known to be an unbiased estimation of the true generalization error [23]. Additionally, the LOO eliminates any subjectivity from the test protocol, hence it can be repeated and exactly the same results will be obtained.

This dataset is truly representative of the real-life situation in the NICU and it allows for a robust estimate of the algorithm performance. In fact, the LOO performance estimated on this dataset was shown to closely match the performance which was independently assessed on a separate large clinical dataset, as reported in [14] and [18].

B. PATIENT INDEPENDENT SEIZURE DETECTION ALGORITHMS (PI-GMM, PI-SVM, PI-FUSION)

The typical patient-independent SDA consists of the following blocks: EEG signal pre-processing, feature extraction, modelling, post-processing and decision making. Three patient independent classification algorithms have been developed employing: the support vector machine classifier (PI-SVM), a Gaussian mixture model based classifier (PI-GMM) and a fusion of the two classifiers (PI-Fusion). More details on these SDAs including the list of extracted features and the probabilistic interpretation of the classifier output can be found in Appendices A-D. These SDAs as described in [14] and [22] are patient independent systems, which have no prior sight of the EEG of the neonate under test. Importantly, these systems have been developed so that every channel of the EEG is processed separately and independent of the other channels, which means that the system is robust to the number and choice of channels.

Both the GMM and SVM classifiers perform an extraction of a compact representation of the training data for each class. For the GMM this is based on the data centroids which are obtained by averaging over the training data. In the case of the SVM however, this is based on a subset of the training data which lies close to the discriminative boundary. While the support vectors are selected from the training data in the context of both classes, the GMM centroids are however class-indifferent, and thus are not optimized to increase the separability of the problem. The previous work on combining SVM and GMM classifiers for neonatal seizure detection showed a significant disparity between classifier decisions, resulting in an agreement of only approximately 50% of the false positives [22].

A simple *blending* of patient-independent GMM and SVM classifiers using the geometric mean is used to provide the patient independent SDA, (PI-FUSION):

$$P_{PI-Fusion} = \sqrt{P_{PI-GMM} P_{PI-SVM}} \quad (1)$$

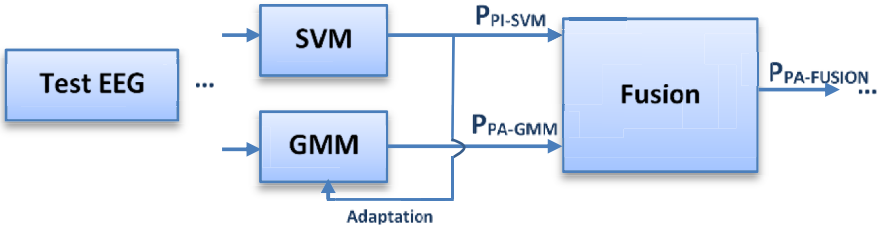

FIGURE 2. The ensemble of patient-adaptive GMM and patient-independent SVM SDAs.

Here P_{PI-GMM} and P_{PI-SVM} are the probabilities of seizure provided by the patient independent GMM and SVM classifiers respectively. The geometric mean combination was found more suitable than the arithmetic mean for fusing classifiers with different probability density functions [20]. In fact, the SVM probabilities usually follow a gamma distribution and the GMM posteriors follow a Gaussian distribution.

C. ORACLE SYSTEMS – PATIENT DEPENDENT SDAS USING PATIENT SPECIFIC CLINICAL LABELS (PD-SVM AND PD-GMM)

The patient dependent seizure detectors, PD-SVM and PD-GMM, were constructed in the same way as their patient-independent alternatives, but with a small portion (a few minutes) of the test patient data (with 'true' clinical neurophysiologist annotations) used in training. These systems, PD-SVM and PD-GMM are referred to as Oracle systems, as they use labels provided by a clinical expert. These systems are used to estimate the theoretical performance improvements that could be obtained if some clinical labels were available for the test patient. In our case, the Oracle systems were not fully patient-dependent as they were still trained with data from other patients (not just with the targeted patient's data). Moreover, no special emphasis was given to the sampling of the targeted patient data – the new data were simply randomly mixed with the existing training data. These resultant systems were then tested on the remaining *unseen* data from that specific test patient – the systems are therefore no longer patient-independent as they have seen samples of seizure and non-seizure activities from the targeted testing patient.

Patient-dependent SDAs are quite popular in the adult population (especially those that are based on intracranial EEG), where the data collected during previous hospital visits are annotated and used to develop patient-specific models for subsequent visits [24], [20]. In fact, it has been shown in [13] that a *fully* patient-dependent neonatal SDA performs much better than a patient-independent one. However, in the neonatal population such systems are impossible as EEG data from the newborn brain is not typically recorded until the baby is born. There is a significant clinical time pressure – the detection system should be functioning and supporting clinical decisions from the moment the EEG electrodes are placed on the newborn's scalp, literally within a few hours of birth.

The Oracle systems used here mimic a scenario where small portion of annotated data of the testing subject is available beforehand: For example when a clinician (a neurophysiologist) who is alerted by an alarm generated by the generic neonatal SDA (SVM or GMM), is able to provide feedback about the true label of this alarm (as shown in Fig. 1). The label (seizure or non-seizure) can then be used to adapt automatically the models with this new information, thus making the models personalised.

D. A SEMI-SUPERVISED LEARNING SCHEME FOR PATIENT ADAPTIVE SDA (PA-GMM AND PA-FUSION)

This section details an alternative technique for the *blending* of the two classifiers where the PI-SVM system is used to automatically label new data, for example from a new unseen patient – these labels can then be used for the on-line adaptation of a GMM based detector (PA-GMM). The final ensemble (PA-FUSION) consists of the fixed patient-independent PI-SVM classifier and the changing patient-adaptive PA-GMM classifier, which are then blended, as shown in Fig. 2.

The SVM paradigm has achieved considerable success in a wide variety of problems in a batch setting where all of the training data is available in advance. Several learning techniques have been developed to facilitate SVM training over very large datasets, however, only a few have been proposed for incremental, online and active learning. Most of active learning techniques, such as adiabatic learning [25], are approximate and require several passes through the data to reach convergence. Although these algorithms allow for training on large datasets that is significantly faster than typical state-of-the-art SVM solvers, they are still incapable of real-time training and allow for little (or no) control over the confidence of the new data during the training procedure. This results in the limited use of these methods in an online setting suitable for real-time medical applications [26].

In contrast to the SVM, several well-established techniques exist to perform online learning for a Gaussian mixture model based classifier. These are widely used in the area of speech processing, for instance to improve the speech recognition accuracy by adapting the generic phonetic models to those of a specific speaker [27], [28].

1) MAP ADAPTATION OF THE GMM MODEL

Given an ordered set of N new feature vectors $\mathbf{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$, a corresponding ordered set of the associated seizure probabilities produced from the PI-SVM, are generated as, $P_{S-SVM} = \left\{P_{S-SVM,1}, P_{S-SVM,2}, \ldots, P_{S-SVM,N}\right\}$. The seizure and non-seizure GMMs, are parameterised by $\theta_S = \left\{\mu_{S,j}, w_{S,j}, \Sigma_{S,j} : \forall j \in \{1, 2, \ldots, M_S\}\right\}$ and $\theta_{NS} = \left\{\mu_{NS,j}, w_{NS,j}, \Sigma_{NS,j} : \forall j \in \{1, 2, \ldots, M_{NS}\}\right\}$, respectively, (see Appendix C). The patient-adaptive models are then developed by the adaptation of the original patient-independent models based on new test patient data. The conventional maximum a-posteriori (MAP) adaptation is used here and consists of the following steps:

a) Compute the occupational likelihood for each feature vector, $\boldsymbol{x}_i$, with respect to the m^{th} Gaussian component of each class model θ_C. The occupational likelihood determines how relevant a particular new feature vector is to the given Gaussian component,

$$P_{C,m}\left(x_i, \theta_C\right) = \frac{w_{C,m} g\left(x_i | \mu_{C,m}, \Sigma_{C,m}\right)}{\sum_{j=1}^{M_C} w_{C,j} g\left(x_i | \mu_{C,j}, \Sigma_{C,j}\right)}. \quad (2)$$

b) Compute the mean of the adaptation data, weighted by the occupational likelihood, over the N new feature vectors, $\mathbf{X}$. For the m^{th} Gaussian component of class C; this yields,

$$E_{C,m}\left(\mathbf{X}\right) = \frac{\sum_{i=1}^{N} P_{C,m}\left(\boldsymbol{x}_i, \theta_C\right) \boldsymbol{x}_i}{\sum_{i=1}^{N} P_{C,m}\left(\boldsymbol{x}_i, \theta_C\right)}. \quad (3)$$

c) Update the new mean of the m^{th} Gaussian component of class C as the weighted average of the original mean and the adaptation data mean,

$$\mu_{C,m} \leftarrow \alpha \mu_{C,m} + (1 - \alpha) E_{C,m}\left(\mathbf{X}\right). \quad (4)$$

In this manner every single mean component of the models for each class are updated based on the weighted average of the original mean and the mean of the adaptation data weighted by the occupational likelihood. Three iterations are used in the MAP adaptation routine in this work.

2) NOVEL MAP ALGORITHM BASED ON THE CONFIDENCE OF THE AUTOMATED LABELS

When applying MAP adaptation, a label is required for the new feature vector to indicate which class specific model should be adapted with these new data. These labels are generated automatically using the PI-SVM based SDA. Such labels could be generated by thresholding the SVM probabilities, for example with a threshold of 0.5. However, the choice of any specific threshold would have a significant effect on the performance of the system and as such would therefore require careful tuning. Moreover, in this approach, the data would be split between the two class specific models – that is, a given chunk of data would be used to adapt either the seizure model or the non-seizure model, depending on its assigned label. Additionally, as seizures are relatively rare events, it is likely that the seizure model would not be adapted at all during the first few hours for a new patient. In order to maximise the power of the data available from a new patient, the MAP adaptation was modified to enable the use of all the new testing data to adapt both the seizure and non-seizure GMM models, simultaneously.

First, the data set of N new feature vectors are grouped into clusters according to their associated PI-SVM probabilistic output. The grouping is performed by partitioning the probability space [0 1] into a set of non-overlapping bins; the k^{th} bin will have a lower limit $\underline{P_k}$ and an upper limit $\bar{P}_k$. Given the ordered set of SVM probabilities $P_{S-SVM} = \left\{P_{SVM,1}, P_{SVM,2}, \ldots, P_{SVM,N}\right\}$, the corresponding ordered set of N new feature vectors, $\mathbf{X} = \{x_1, x_2, \ldots, x_N\}$ is then clustered for the seizure model using the rule, $i \in I_{S,k}$ if $\underline{P_k} \leq P_{SVM,i} < \bar{P}_k$, where $I_{S,k}$ is the indicial set for the k^{th} cluster, for seizure model adaptation. The N new feature vectors would also be clustered for the non-seizure model using the complementary rule, $i \in I_{NS,k}$ if $\underline{P_k} \leq \left(1 - P_{SVM,i}\right) < \bar{P}_k$, where $I_{NS,k}$ is the indicial set for the k^{th} cluster, for the non-seizure model adaptation. These clusters represent different confidences of being relevant to a chosen class.

The MAP adaptation algorithm can now be reformulated as a weighted combination of the statistics of each of the K groups, with the weight-set, $\{\beta_1, \beta_2, \ldots, \beta_K\}$,

$$\mu_{C,m} \leftarrow \alpha \mu_{C,m} + \sum_{k=1}^{K} \beta_k \left(\frac{\sum_{i \in I_{C,k}} P_{C,m}\left(\boldsymbol{x}_i, \theta_C\right) \boldsymbol{x}_i}{\sum_{i \in I_{C,k}} P_{C,m}\left(\boldsymbol{x}_i, \theta_C\right)} \right),$$

where

$$C \in \{S, NS\}. \quad (5)$$

The weights sum up to 1; the weight of the original data, $\alpha = 1 - \sum_{k=1}^{K} \beta_k$, determines how aggressive the adaptation will be on the new data.

Intuitively, the cluster with a higher $\bar{P}_k$ should have a larger gain, β_k. The four basic and intuitive weighting schedules shown in Fig. 3 (a) are investigated. The group weights, $\{\beta_1, \beta_2, \ldots \beta_K\}$, for both classes follow either of a simple straight line, a sigmoid, or two exponential functions each with a different decay rate. In all cases the weights increase monotonically with confidence. In contrast to the linear function which gives linearly decaying weights to the groups as confidence reduces, the sigmoid function nonlinearly emphasises the high-confidence groups and attenuates the low confidence groups. The two exponential functions are more conservative than the linear and sigmoid functions, in that they allow for the adaptation of the class-specific models to be focussed only on new data where the label confidence is high.

The number of confidence-based data groups is first chosen and the continuous probability space [0, 1] in Fig. 3 (a) is then partitioned to provide K clusters. Fig. 3 (b) shows how for example a sigmoid is sampled to provide the weights for 5 clusters. Every point in the k^{th} cluster is assigned the same central weight h_k; after sampling these weights are then

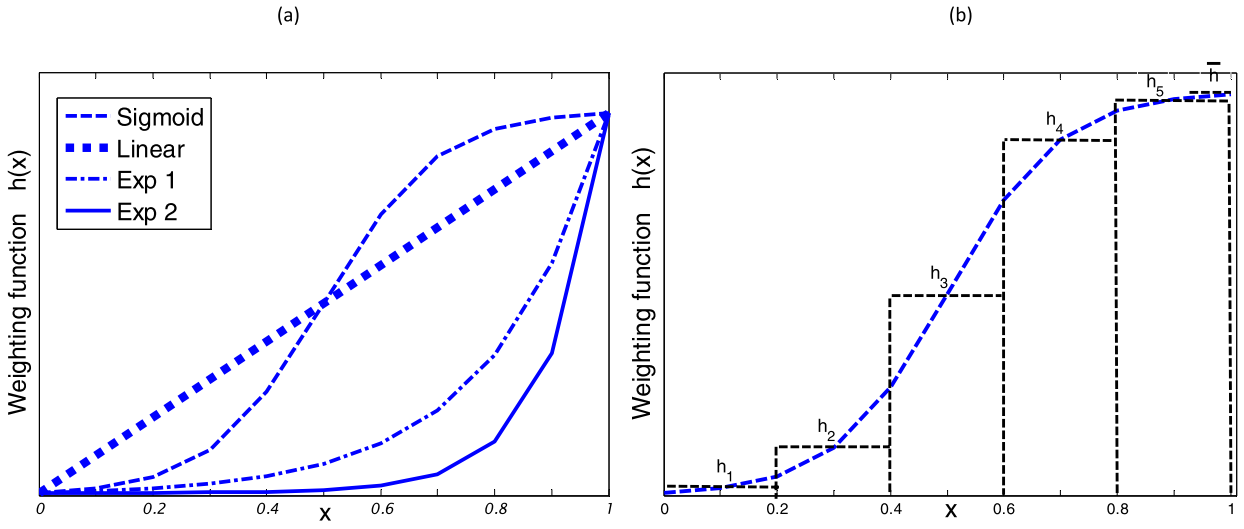

FIGURE 3. **(a) Four different weight functions; here x = PSVM for providing weights for updates to seizure class GMM, x = (1-PSVM) for updating non-seizure class GMM. (b) An example of a sampled sigmoid weighting using 5 clusters.**

normalized to provide a partition of unity,

$$\beta_k = \frac{h_k}{\bar{h} + \sum_{k=1}^{K} h_k}. \tag{6}$$

The weight of the original data, α, is always the largest and effectively places more confidence on the original model for which the training data (with clinical labels) are more certain.

The resultant patient-adaptive GMM system (PA-GMM) is then blended with the patient-independent SVM classifier (PI-SVM) using the Eq. 1 to form the final ensemble (PA-FUSION).

E. PERFORMANCE EVALUATION

The system performance is measured as the average area under the receiver operating characteristic curve (AUC) [29]. The AUC is calculated by plotting the sensitivity vs specificity values computed over the probabilistic output produced for every epoch of EEG. The area under the ROC curve is an effective way of comparing the performance of different systems - a random discrimination will give an area of 0.5 under the curve while perfect discrimination between classes will give unity area under the ROC curve. Additionally, the AUC90 is reported where the area under the curve is computed for a specificity larger than 90%. The AUC90 is more reflective of the potential clinical scenario as it focuses and quantifies the performance in the area with very low false detection rates.

Fig. 4 shows an example of the ROC curve for a typical SDA from which AUC and AUC90 can be computed. The mean AUC area across all patients is reported in this study. The ROC area is related to the Wilcoxon test of significance [30]. This relationship can be used to derive statistical properties of the ROC area such as its standard error and to calculate the statistical significance in the performance of two algorithms (ROC areas) evaluated on the same data; this takes into account the correlation of the two ROC curves [14], [30], [36]. The details of the statistical test can be found in Appendix E.

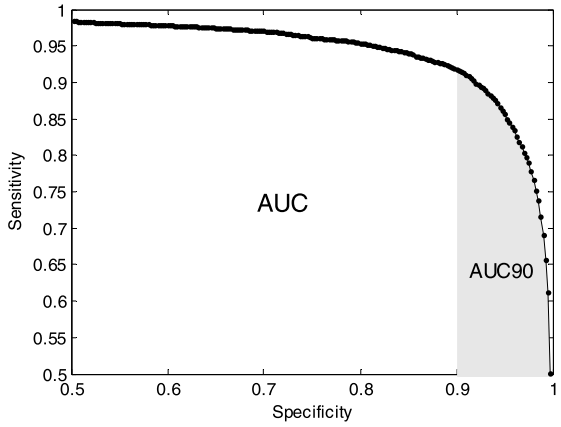

FIGURE 4. **Performance of a SDA - measured as the area under the ROC curve.**

TABLE 1. **Performance of the neonatal seizure detection systems.**

		AUC (%)	AUC90 (%)
Baseline	PI-GMM	95.70	78.40
	PI-SVM	96.50	82.60
	PI-FUSION	96.62	82.61
Oracle	PD-GMM	97.51	86.33
	PD-SVM	97.17	85.80
Adaptive	PA-GMM	96.91	82.60
	PA-FUSION	**97.03**	**84.10**

III. EXPERIMENTAL RESULTS

Table 1 summarises the performance scores of the various systems: Baseline patient independent systems

IEEE Journal of Translational Engineering in Health and Medicine

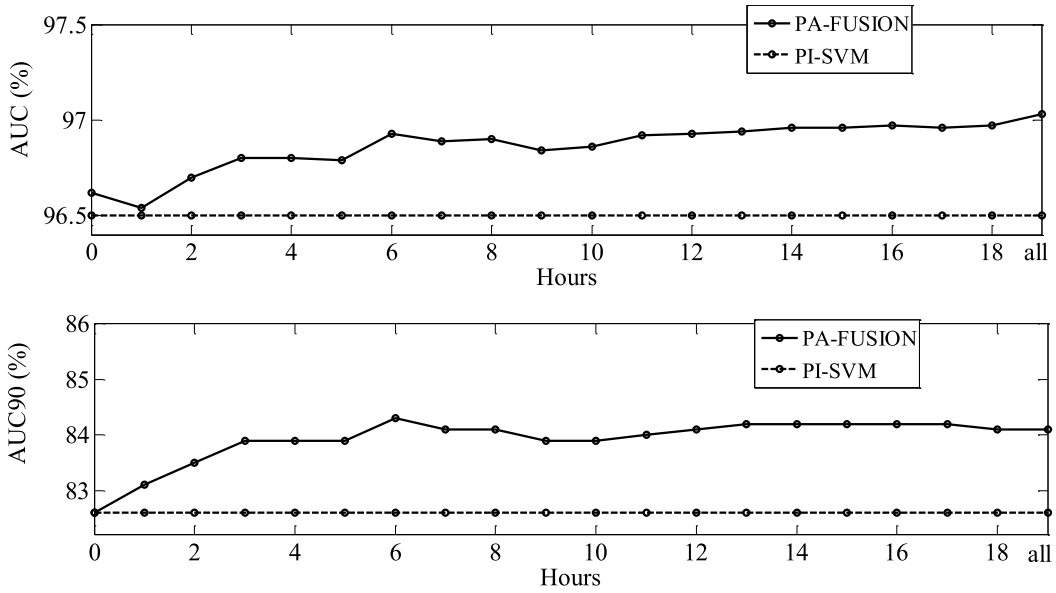

FIGURE 5. Performance of the baseline SVM and adaptive fusion system measured as AUC (top) and AUC90 (bottom) as a function of time.

(PI-SVM, PI-GMM and PI-FUSION), the Oracle systems using some clinical labels for each test baby (PD-SVM, PD-GMM) and the Adaptive systems in which automated labels are generated, (PA-GMM and PA-FUSION).

The SVM patient-independent system (PI-SVM) on this dataset yields a performance of 96.50% and 82.6% for AUC and AUC90, respectively. The GMM patient-independent system (PI-GMM) on this dataset yields a performance of 95.70% and 78.49% for AUC and AUC90, respectively. The performance of a simple blending of GMM and SVM (PI-Fusion) results in a performance of 96.62% and 82.65% for AUC and AUC90. As can be seen, the performance of the ensemble (PI-FUSION) is slightly improved in comparison with the best single patient-independent classifier (PI-SVM), which has an AUC of 96.50%.

The performance of the Oracle patient-dependent system, PD-SVM, is 97.17% and 85.80%, for AUC and AUC90, respectively. The performance of the Oracle patient-dependent system, PD-GMM is 97.51% and 86.33% for AUC and AUC90, respectively.

The adaptive GMM system (PA-GMM) provides a performance which is an improvement over its patient-independent GMM counterpart (PI-GMM) – 96.91% vs 95.70% for AUC, and 82.6% vs 78.4% for AUC90. In fact, it also outperforms the best baseline patient-independent SVM system (PI-SVM), 96.91% vs 96.50% for AUC, with the same performance as measured by AUC90.

Fig. 5 shows how the mean AUC and AUC90 (determined over all the unseen records in the LOO performance assessment routine) evolve with training time, as compared to the baseline PI-SVM performance. In this experiment, the data from each unseen baby within the LOO validation scheme is split into one hour segments. The PA-GMM is then adapted sequentially on each hour of data, until eventually all the data was used. The labels were produced automatically using the PI-SVM – an exponential weighting function (exp2) was used with 10 groups. The performance of the fused classifier PA-FUSION was reported after each hour of adaptation – this was evaluated over the whole recording from the beginning, to allow for a fair comparison with the baseline performance. This procedure was repeated for each baby within the LOO scheme, where for each unseen test baby the GMM was re-initialised to the PI-GMM. The last point on the curves in Fig. 5 indicates the offline performance of the fused classifier, that is, it provides the performance of a retrospective corrected viewing of the unseen EEG recording. Similarly, the first point on the curves in Fig. 5 represents the performance of the PI-FUSION system which is simply the geometric mean combination of the baseline PI-SVM and PI-GMM, when adaptation has not yet been performed. This evaluation strategy mimics the real-life operation of the neonatal SDA – the system runs in real-time but it may re-adjust or correct its previous decisions based on the evidence observed.

Table 2 shows the effect the choice of the weights on the adaptive systems. Here four different weight profiles were considered: linear, sigmoidal, an exponential decay, and a more aggressive exponential decay curve. Table 3 presents the performance of the proposed patient adaptive systems for different numbers of data clusters.

Fig. 6 shows the relative improvement in AUC90 for each subject in the dataset between PA-FUSION and PI-FUSION, to illustrate the effect of adaptation. The relative improvement is calculated as (AUC90PA-FUSION – AUC90PI-FUSION)/(1 – AUC90PI-FUSION). The statistically significant difference at α set to 1% as in [14] is indicated with the asterisk.

TABLE 2. Performance with different adaptation weighting functions.

		Function	AUC (%)	AUC90 (%)
Adaptive	PA-GMM	Sigmoid	96.70	81.6
		Linear	96.79	82.3
		Exp 1	96.85	**82.9**
		Exp 2	**96.91**	82.6
	PA-FUSION	Sigmoid	96.97	83.8
		Linear	96.99	84.0
		Exp 1	97.02	**84.2**
		Exp 2	**97.03**	84.1

TABLE 3. Performance with different number of groups.

		# Groups	AUC (%)	AUC90 (%)
Adaptive	PA-GMM	8	**96.93**	82.5
		10	96.91	82.6
		15	96.90	82.7
	PA-FUSION	8	**97.03**	84.0
		10	**97.03**	84.0
		15	**97.03**	**84.1**

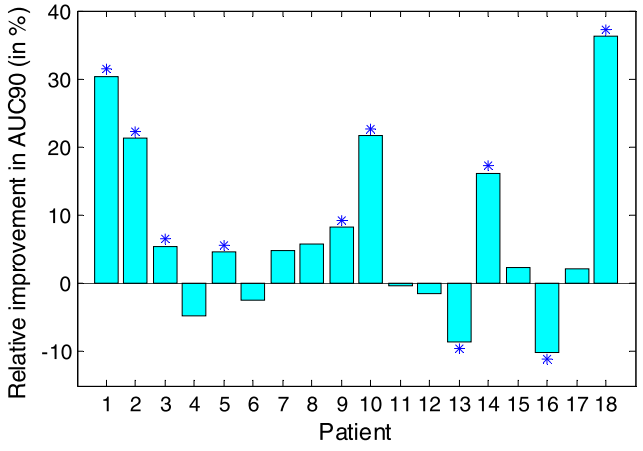

FIGURE 6. Relative improvement in AUC90 between the PA-FUSION and PI-FUSION systems. '*' indicates statistical significance at á set to 1%.

IV. DISCUSSION

A. PERFORMANCE OF THE ORACLE SYSTEMS (PD-SVM AND PD-GMM)

This work explores the unsupervised use of patient specific test data to improve the developed models. It can be seen that both PD-SVM and PD-GMM yield performance improvements as compared with their patient-independent counterparts. The SVM system improved its AUC from 96.50% to 97.17% and its AUC90 from 82.6% to 85.80% The GMM system improved its AUC from 95.70% to 97.51% and its AUC90 from 78.49% to 86.33%. Comparing the improvements and the absolute performances of the Oracle systems from Table 1, it can be seen that PD-GMM is more sensitive to the new data and better exploits even small amounts of the targeted patient data. The GMM based classifier is also much easier to adapt because of the availability of the well-established model adaptation routines – this is highlighted by the performance improvement achieved for the GMM. As a comparison, however, the baseline SVM patient-independent SDA achieved a better performance (96.50% vs 95.70% for AUC, and 82.6% vs 78.49% for AUC90) when compared with the patient independent GMM – this shows that for our experiments that the SVM provided better performance than the GMM when dealing with batch data; this is also confirmed in [22].

An important conclusion that can be extracted from the performance of the Oracle systems is the upper bound on the performance of automatic adaptive systems. In fact, what the automatic adaptive SDA does is to create a system-generated label to avoid consulting a clinician (or to keep functioning in the absence of a neurophysiologist) and to use this label to adapt the models. As no SDA is error-free, the use of the true label can be seen as an estimate of the maximum achievable performance given the chosen methodology.

B. PERFORMANCE OF THE PATIENT-ADAPTIVE SYSTEMS (PA-GMM AND PA-FUSION)

The results of the proposed adaptive SDAs demonstrate the ability of the proposed method to capture the test patient specifics by the automatic, unsupervised (in the sense that there is no human involved), on-the-fly adaptation of the GMM models using well-established adaptation routines as a core. The combination of PI-SVM and PA-GMM as an adaptive fusion system (PA-FUSION) provided comparable and improved results in AUC and AUC90 values in comparison with the GMM adaptive system alone (PA-GMM), 97.03% vs 96.91% for AUC, and 84.1% vs 82.6% for AUC90. This indicates that the PA-GMM system still carries complementary information to the PI-SVM system which is exploited with the geometric mean fusion – even though it was adapted based on the automatic labels provided by the PI-SVM classifier.

The Oracle systems achieve the best possible results. The comparison of the adaptive systems with the Oracle indicates that the adaptive SDA performance is close to that of the human-supervised adaptation (Oracle). We can also conclude that the use of a relatively small amount of manually-annotated data can lead to better performance than unsupervised adaptation with a lot of data – this is not unexpected given that the decisions produced by the PI-SVM classifier contain errors; these errors inevitably affect the model purity.

C. HOW THE PERFORMANCE OF PATIENT ADAPTIVE SYSTEMS (PA-GMM AND PA-FUSION) DEPEND ON THE DURATION OF ADAPTATION TO A SPECIFIC PATIENT

From Fig. 5 it can be appreciated that the performance of the adaptive system increases rapidly during the first 6 hours of recording with little or no additional performance improvement after this time. Since the adaptation is driven by the PI-SVM system, the PA-GMM adaptive system gradually changes its nonlinear decision boundaries to approximate

IEEE Journal of Translational Engineering in Health and Medicine

those of the SVM system - thus improving the GMM performance for data for which the SVM was confident. However, it is essential that the GMM is adapted to effectively fill in the gaps in its performance, rather than to be over-trained to fully mimic the PI-SVM. Therefore, as shown in Fig. 5, the similarity between the outputs of the two classifiers increases with increasing data which then leads to a slow-down in the progress of the adaptive fusion system (PA-FUSION). The adaptation data allow the GMM adaptive system to learn the specifics of the test patient at the cost of incorporating into its models the errors that come with the imperfect SVM hypotheses. At the same time, the amount of the adaptation data for the GMM adaptive system can be seen as a trade-off between learning new information (and thus being more accurate under the above-mentioned constraints) and being different (and thus complementary) to the SVM baseline. It can be seen that the in-built intrinsic difference between the two classifiers such as the generative GMM and the discriminative SVM allows the adaptive fusion system to benefit from adaptation and to maintain a stable performance even until the end of the recording.

Our previous work has demonstrated that the performance of the PI-SVM and PI-GMM systems had a standard deviation of AUC of $\sim$2% across all iterations of the LOO validation [22]. The performance of the presented patient-adaptive method does not therefore depend on the group of patients used in the training of the PI-SVM but rather on the specifics of the testing patient – the amount of seizure data present within the first few hours of recording.

D. THE INFLUENCE OF THE WEIGHT FUNCTION AND THE NUMBER OF DATA GROUPS ON THE PERFORMANCE OF PA-GMM AND PA-FUSION SYSTEMS

The weights in Eq. 5 and Eq. 6 control the trade-off between the gain of learning new patient-specific information and the cost of introducing noise into the models – non-seizure characteristics to the seizure model and seizure characteristics to the non-seizure model. It can be seen that the GMM adaptive systems (PA-GMM) with any of these weighting profiles provided better performance than the PI-GMM system (results in Table 1, AUC = 95.70%). The aggressive exponential decay weighting (Exp2) provides the best trade-off (AUC = 97.03% as compared to 95.7%). This is a conservative weighting function – the faster the decay, the lower the influence of low-confidence groups in the adaptation framework. For the number of clusters, the performance was observed to be stable in the tested range (5 – 25 clusters).

It is worth noting that the parameters of the system were not tuned to reach the best possible performance. The aim of this study was to demonstrate the potential of the unsupervised on-line adaptation for improved performance in neonatal SDA through the personalisation of detection algorithms. Further performance improvements should be possible, if instead of sampling common generic weighting functions, the weights could instead be estimated on the training data using maximum likelihood optimisation, for example.

E. TRANSLATIONAL RELEVANCE AND STATISTICAL SIGNIFICANCE

From Fig. 6 the comparison between patient-adaptive and patient-independent ensemble systems can be performed with the test of statistical significance (Appendix E). Both systems represent a fusion of PI-SVM and either PI-GMM or PA-GMM. The significance level to reject the null hypothesis was set to 1% ($\alpha = 0.01$). It can be seen that the adaptation improves the AUC90 significantly in 8 patients and decreases the performance in 2 patients. In 4 out of 8 patients with the improved AUC90, this increase was above 20%. In the remaining 8 patients, the performance, as measured in AUC90 with the chosen cut-off point, does not change significantly. The average relative improvement in AUC90 in those patients whose AUC90 changed significantly is approximately 10%.

Event based metrics provide a better measure of the clinical benefits of the adaptive fusion system. After 7 hours of unsupervised adaptation for each unseen patient in the LOO scheme, the average seizure detection rate (over all unseen babies) improved from 63% to 70%, while keeping a fixed false detection rate of 0.2 FD/h (on average 1 false detection every 5 hours). This corresponds to an additional 90 detected seizures detected over the whole database. Alternatively, focusing on the reduction of the FD/h rate, while maintaining a seizure detection rate at 70%, the number of false detections per hour was halved, from 0.4 FD/h to 0.2 FD/h.

An example of the real-time functioning of the developed algorithm is shown in Fig. 7, where the probabilistic output of the PA-FUSION SDA is contrasted with the probabilistic output of the baseline PI-SVM SDA for 1 hour of EEG with the superimposed clinical annotations. It can be seen that both seizures in the example result in a higher probability output from the PA-Fusion system, whereas the probabilistic output for the non-seizure EEG in-between is attenuated. This improvement comes from learning patient-specific information in an unsupervised way which makes the system more robust to inter-patient variability and allows it to focus the learning process on the difference between seizure and non-seizure characteristics.

It is important to realise that from an event detection and hence clinical point of view, this technique provides performance improvements. If a decision threshold of 0.5 was utilised in Fig. 7, then both systems would detect the two seizure events shown; however there would be 4 false alarms with the PI-SVM approach which would be reduced to 1 false alarm with the proposed patient adaptive approach. If the threshold was increased to 0.6, the PI-SVM approach would miss one seizure event with the PA-FUSION approach still detecting the 2 events. There would now only be one false detection for the PI-SVM detector; this was improved to zero false detections using the proposed PA-Fusion detector.

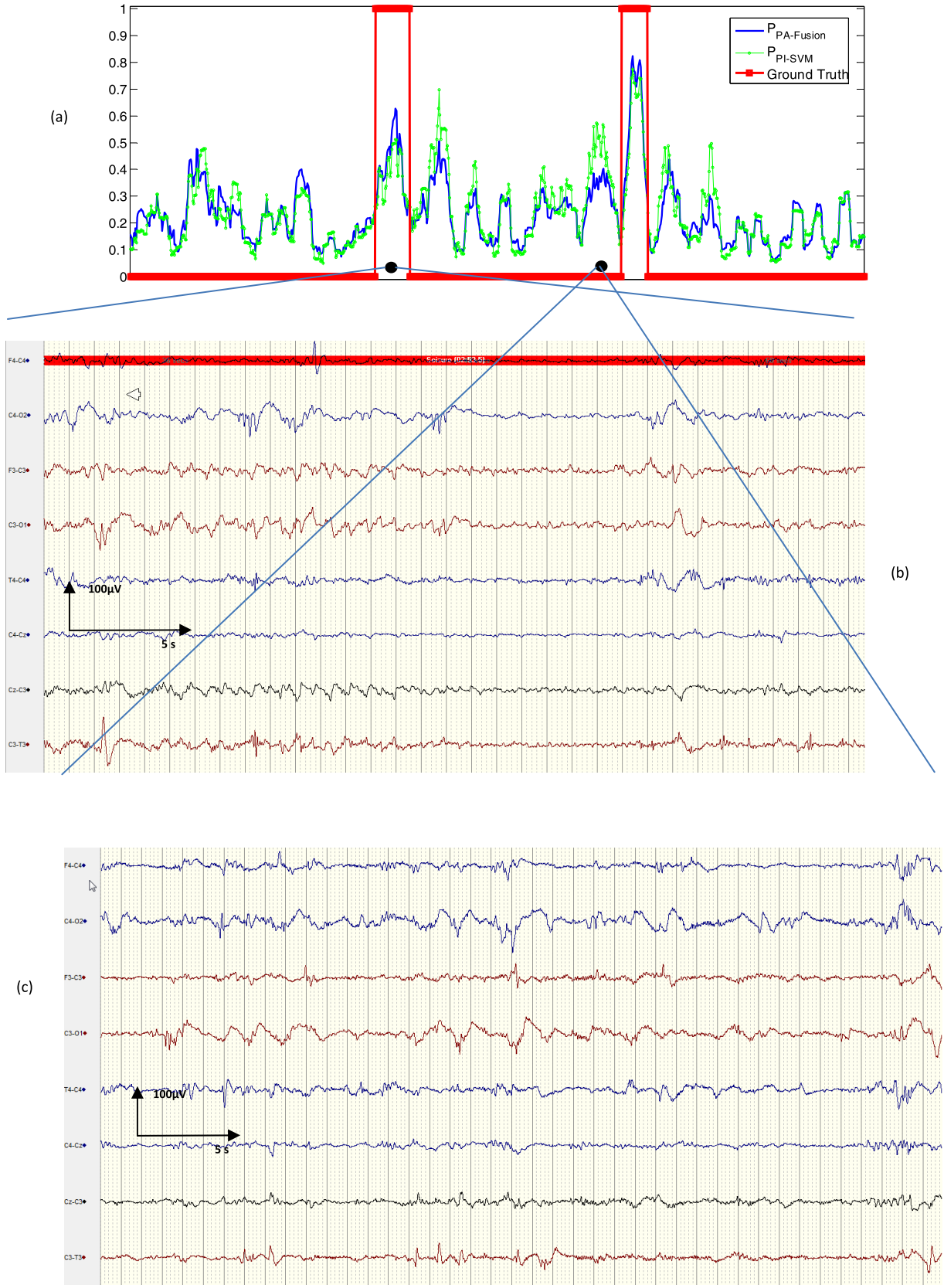

FIGURE 7. Real-time functioning of the developed personalised SDA algorithm. A) Probabilistic output of patient-adaptive and patient-independent SDAs with superimposed ground truth. B) 30s of seizure activity detected with higher probability with adaptive fusion system. C) 30s of non-seizure activity detected with lower probability.

IEEE Journal of Translational Engineering in Health and Medicine

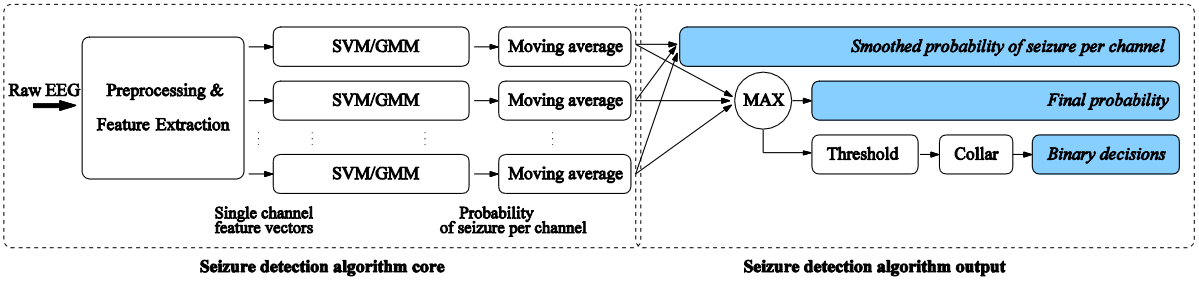

FIGURE 8. Neonatal seizure detection system diagram.

TABLE 4. Extracted features of neonatal EEG.

Domains	Feature List
Frequency	• Total power (0-12Hz) • Peak frequency of spectrum • Spectral edge frequency (80%, 90%, 95%) • Power in 2Hz width subbands (0-2Hz, 1-3Hz, ...10-12Hz) • Normalised power in subbands • Wavelet energy
Time	• Non-linear line length • Number of maxima and minima • Root mean squared amplitude • Hjorth parameters • Zero crossings (raw epoch, Δ and ΔΔ) • Autoregressive modeling error (model order 1-9) • Skewness • Kurtosis • Nonlinear energy • Variance (Δ and ΔΔ)
Information theory	• Shannon entropy • Singular value decomposition entropy • Fisher information • Spectral entropy

V. CONCLUSIONS

This study has contributed to the implementation of personalised healthcare in the area of seizure detection in the newborn. A combination of patient adaptive generative and patient independent discriminative classifiers has led to an improvement in the detection of neonatal seizures over the course of long EEG recordings, as validated on a long unedited EEG dataset. More accurate detection comes from both the different nature of the classification approaches and the real-time incorporation of patient-specific data. To the best of our knowledge, this is the first study to propose the use of online adaptation to build a personalized diagnostic system for detection of neonatal seizures.

APPENDIX A
EEG PRE-PROCESSING AND FEATURE EXTRACTION

The EEG is filtered with a band-pass zero-phase filter [0.5-13Hz] and down-sampled from 256 to 32 Hz. A 55-dimensional feature vector is extracted from an 8-second long single-channel EEG epoch, with 50% overlap. The feature set considered have been shown to be useful for neonatal seizure detection in a number of papers from different groups [14], [31] and can capture temporal, frequency, structural and energy information about neonatal seizures. The features extracted are listed in Table 4. More details can be found in [14]. The feature vectors are normalized by subtracting the mean and dividing by the standard deviation to assure commensurability of various features. These 110 normalisation constants (55 means and 55 standard deviations) are estimated on the training data and applied to the test data.

APPENDIX B
AN SVM BASED PATIENT INDEPENDENT CLASSIFIER

The SVM-based patient independent neonatal SDA is shown in Fig. 8 and is described in detail in [5] and [14]. For a test vector $\boldsymbol{x} \in R^d$, and $y \in \{-1, +1\}$, the output decision of a support vector machine is given by:

$$y(\boldsymbol{x}) = \text{sgn}(f(x))$$

where

$$f(x) = b + \sum\nolimits_{j \in Q_{SVM}} \alpha_j y_j K\left(\boldsymbol{x}, \tilde{\boldsymbol{x}}_j\right) \tag{7}$$

Here Q_{SVM} is the indicial set of the retained vectors from the input training data which are called support vectors. The j^{th} support vector, $\tilde{\boldsymbol{x}}_j$, has an associated training label, y_j, and a non-zero weight α_j; b determines the offset of the separating hyperplane from the origin.

The Gaussian kernel is used:

$$K\left(\boldsymbol{x}, \tilde{\boldsymbol{x}}_j\right) = exp\left(-\frac{1}{2\sigma^2}\left(\boldsymbol{x} - \tilde{\boldsymbol{x}}_j\right)^T\left(\boldsymbol{x} - \tilde{\boldsymbol{x}}_j\right)\right) \tag{8}$$

The SVM model was trained on data with per-channel seizure and non-seizure annotations, thus resulting in a single generic SVM classifier that can be applied to any EEG channel from any patient. This classifier was trained to maximise the margin, with softening of this margin included to allow a balance between decision errors and over-fitting. In the dual form, the training algorithm can be stated as the following optimisation problem over the training set consisting of N_T input feature vectors, $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{N_T}\}$ with associated

labels $\{y_1, y_2, \ldots, y_{N_T}\}$ where $y_i \in \{-1, +1\}$.

$$\max_{\alpha} \left(\sum_{i=1}^{N_T} \alpha_i - \frac{1}{2} \sum_{i=1}^{N_T} \sum_{j=1}^{N_T} \alpha_i \alpha_j y_i y_j K\left(\boldsymbol{x}, \tilde{\boldsymbol{x}}_j\right) \right)$$
$$s.t. \sum_{i=1}^{N_T} \alpha_i y_i = 0 \quad \text{and } 0 \le \alpha_i \le C, \quad i = 1, 2, \ldots, N_T \tag{9}$$

The regularisation constant C helps to control the degree of over-fitting.

Cross-validation on the training data is used to select suitable model parameters for the regularisation constant C and the Gaussian kernel width σ. During the testing stage, the features extracted from each EEG channel of multi-channel EEG are fed into the trained model so that the SVM output is generated separately for each EEG channel. The output of each SVM is converted to posterior probabilities as explained in Appendix D and then smoothed using a 15$^{\text{th}}$ order moving average filter which corresponds to a span of ~1 minute of EEG. Depending on the user interface chosen, the output of the system can be either per-channel probabilities, or a single probability trace (by taking the maximum of probabilities across channels), or binary decisions which are produced by applying a threshold to the final probability followed by a collar to compensate for the delay introduced by the moving average filter.

APPENDIX C
A GENERATIVE APPROACH – USING THE GAUSSIAN MIXTURE MODEL

The GMM is a generative approach in which the class-specific probability density functions over the d dimensional feature space are first modelled using a weighted sum of M_C Gaussian components, for each class C,

$$\begin{aligned} p_c\left(\boldsymbol{x}|\theta_C\right) &= \sum_{j=1}^{M_C} \frac{w_{C,j}}{\sqrt{(2\pi)^d \left|\Sigma_{C,j}\right|}} \\ &\quad \times exp\left(-\frac{1}{2} \left(\boldsymbol{x} - \mu_{C,j}\right)^T \Sigma_{C,j}^{-1} \left(\boldsymbol{x} - \mu_{C,j}\right) \right) \\ &= w_{C,j} g\left(\boldsymbol{x}|\mu_{C,j}, \Sigma_{C,j}\right) \end{aligned} \tag{10}$$

In the neonatal seizure detection problem presented here it is assumed that there are only two classes, $C \in \{non-seizure\ (NS),\ seizure\ (S)\}$. Again $\boldsymbol{x} \in R^d$ is a feature vector, the mixture weights for class C are $w_{C,j}$, $j = 1, 2, \ldots, M_C$. The j^{th} Gaussian component for class C is a d dimensional Gaussian function with mean vector $\mu_{C,j}$ and covariance matrix $\Sigma_{C,j}$. The set of M_c mean vectors weights and covariance matrices then form the parameter set θ_C for class C.

In contrast to [22], the training procedure of the seizure and non-seizure models is different in this study. First, a so-called Universal Background Model (GMM-UBM) [32] is constructed by training on all the available training data: seizure, non-seizure, artefactual. This step does not require any annotations and can benefit from the vast amount of data that is available which do not have annotations. The concept of the GMM-UBM has evolved from the speaker identification area where it was used to create a general model of human speech [32] from which a more accurate model of a targeted speaker could be derived. Similarly, in this case, the seizure and non-seizure models are derived from the GMM-UBM using Maximum a Posteriori (MAP) adaptation based on labelled class-specific data [33]. The main advantage of building seizure and non-seizure models through the GMM-UBM is that the class models inherit from the GMM-UBM the wide diversity of signal characteristics across all possible EEG states that would not have been modelled with the direct training of individual models. In effect, this controls the response of the models to 'other' EEG activity. If this activity is equally distanced from both seizure and non-seizure models its effect will be the same and hence it will not contribute to the decision making.

Principle Component Analysis (PCA) is first applied on the features to de-correlate them and reduce the dimensionality. This allows a diagonal covariance matrix to be used in the GMM. In the PCA transformation, 99% of the cumulative energy of the original space is retained whilst reducing the original 55 dimensional feature space to typically 20-25 dimensions (this depends on the training set used in the leave-one-out scheme). The 64 Gaussian UBM model was trained using the Expectation Maximisation (EM) algorithm, [33], [34]. Class specific models were then trained, one for seizure and one for non-seizure.

In the testing stage, the feature vectors of the test EEG data are scored against the seizure and non-seizure models and the output likelihood values are then converted into probabilities using Bayes' theorem. If equal priors are assumed, then similar to Eq. 7, the decision boundary for the GMM can be expressed as,

$$y\left(\boldsymbol{x}\right) = \text{sgn}\left(f\left(x\right)\right)$$

where

$$f\left(x\right) = \ln P_S\left(\boldsymbol{x}|\theta_S\right) - \ln P_{NS}\left(\boldsymbol{x}|\theta_{NS}\right) \tag{11}$$

The same post-processing stage, as is utilised for the SVM system is used with the GMM classifier as shown in Fig. 8.

APPENDIX D
A PROBABILISTIC OUTPUT FOR SVM AND GMM CLASSIFIERS

To determine the class membership of each input vector both classification algorithms calculate a weighted Gaussian distance (a similarity measure) from a test feature vector to each class, represented by either the Gaussian centroids for GMM or the support vectors for SVM. The output of the GMM and SVM can be converted to a probability of seizure using a sigmoid function, in the form:

$$P\left(S|f\left(\boldsymbol{x}\right)\right) = \frac{1}{1 + exp\left(\lambda f\left(\boldsymbol{x}\right) + \varepsilon\right)} \tag{12}$$

IEEE Journal of Translational Engineering in Health and Medicine

where $f(\boldsymbol{x})$ is provided for the SVM and the GMM classifiers using Eq. 7 and Eq. 11, respectively. For the GMM classifier, Bayes theorem provides $\lambda = -1$ and $\varepsilon = 0$. For the SVM classifier, the parameters λ and ε are trained using gradient descent over a subset of the training data [35].

APPENDIX E
A COMPARISON OF AUCS FOR TWO ALGORITHMS

The ROC area is related to the Wilcoxon test of significance [14], [30], [36]. This relationship can be used to derive statistical properties of the ROC area such as its standard error (SE):

$$SE(\gamma) = \sqrt{\frac{\gamma(1-\gamma)+(n_A-1)(Q_1-\gamma^2)+(n_N-1)(Q_2-\gamma^2)}{n_A n_N}} \quad (13)$$

where γ is the ROC area, $Q_1 = \gamma/(2-\gamma)$ and $Q_2 = 2\gamma^2/(1+\gamma)$, n_A and n_N are the numbers of seizure (abnormal) and non-seizure (normal) epochs. To calculate the statistical significance of a difference of the performance (AUCs) of the two algorithms evaluated on the same data, we compute the z statistic by taking into account the correlation of the two ROC curves [36]:

$$z = \frac{\gamma_1 - \gamma_2}{\sqrt{SE(\gamma_1)^2 + SE(\gamma_2)^2 - 2rSE(\gamma_1)SE(\gamma_2)}} \quad (14)$$

where γ_1 and γ_2 refer to the observed areas associated with algorithm 1 and 2, respectively. Here, r represents the estimated correlation between the two ROC curves. The resultant p values of the two-tailed test are reported and values less than 0.01 are considered significant.

Acknowledgment

These financial bodies had no role in the collection, analysis and interpretation of data or the writing of this manuscript. Conflict of interest: None of the authors have potential conflicts of interest to be disclosed.

REFERENCES

[1] A. Harvey *et al.*, "The future of technologies for personalised medicine," *New Biotechnol.*, vol. 29, no. 6, pp. 625–633, 2012.
[2] C. Guger, H. Ramoser, and G. Pfurtscheller, "Real-time EEG analysis with subject-specific spatial patterns for a brain-computer interface (BCI)," *IEEE Trans. Rehabil. Eng.*, vol. 8, no. 4, pp. 447–456, Dec. 2000.
[3] A. Shoeb, H. Edwards, J. Connolly, B. Bourgeois, T. Treves, and J. Guttag, "Patient-specific seizure onset detection," in *Proc. 26th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, vol. 5. Sep. 2004, pp. 483–498.
[4] M. Whirl-Carrillo *et al.*, "Pharmacogenomics knowledge for personalized medicine," *Clin. Pharmacol. Therapeutics*, vol. 92, pp. 414–417, Oct. 2012.
[5] A. Temko, E. Thomas, W. Marnane, G. Lightbody, and G. Boylan, "EEG-based neonatal seizure detection with support vector machines," *Clin. Neurophysiol.*, vol. 122, no. 3, pp. 464–473, 2011.
[6] J. Mitra *et al.*, "A multi-stage system for the automated detection of epileptic seizures in neonatal EEG," *J. Clin. Neurophysiol.*, vol. 26, no. 4, pp. 218–226, 2009.
[7] P. Celka and P. Colditz, "A computer-aided detection of EEG seizures in infants: A singular-spectrum approach and performance comparison," *IEEE Trans. Biomed. Eng.*, vol. 49, no. 5, pp. 455–462, May 2002.
[8] A. Temko, G. Lightbody, E. Thomas, G. Boylan, and W. Marnane, "Instantaneous measure of EEG channel importance for improved patient-adaptive neonatal seizure detection," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 3, pp. 717–727, Mar. 2012.
[9] W. Deburchgraeve *et al.*, "Automated neonatal seizure detection mimicking a human observer reading EEG," *Clin. Neurophysiol.*, vol. 119, no. 11, pp. 2447–2454, 2008.
[10] G. M. K. Ntonfo, G. Ferrari, R. Raheli, and F. Pisani, "Low-complexity image processing for real-time detection of neonatal clonic seizures," *IEEE Trans. Inf. Technol. Biomed.*, vol. 16, no. 3, pp. 375–382, May 2012.
[11] L. Rankine, N. Stevenson, M. Mesbah, and B. Boashash, "A nonstationary model of newborn EEG," *IEEE Trans. Biomed. Eng.*, vol. 54, no. 1, pp. 19–28, Jan. 2007.
[12] J. Altenburg, R. J. Vermeulen, R. L. M. Strijers, W. P. F. Fetter, and C. J. Stam, "Seizure detection in the neonatal EEG with synchronization likelihood," *Clin. Neurophysiol.*, vol. 114, no. 1, pp. 50–55, 2003.
[13] A. Aarabi, R. Grebe, and F. Wallois, "A multistage knowledge-based system for EEG seizure detection in newborn infants," *Clin. Neurophysiol.*, vol. 118, no. 12, pp. 2781–2797, 2007.
[14] A. Temko, G. Boylan, W. Marnane, and G. Lightbody, "Robust neonatal EEG seizure detection through adaptive background modelling," *Int. J. Neural Syst.*, vol. 23, no. 4, p. 1350018, 2013.
[15] S. Vanhatalo, "Development of neonatal seizure detectors: An elusive target and stretching measuring tapes," *Clin. Neurophysiol.*, vol. 122, no. 13, pp. 435–437, 2011.
[16] A. Temko, W. Marnane, G. Boylan, and G. Lightbody, "Clinical implementation of a neonatal seizure detection algorithm," *Decision Support Syst.*, vol. 70, pp. 86–96, Feb. 2015.
[17] S. Mathieson *et al.*, "In-depth performance analysis of an EEG based neonatal seizure detection algorithm," *Clin. Neurophysiol.*, vol. 127, no. 5, pp. 2246–2256, 2016.
[18] S. R. Mathieson *et al.*, "Validation of an automated seizure detection algorithm for term neonates," *Clin. Neurophysiol.*, vol. 127, no. 1, pp. 156–168, 2016.
[19] L. I. Kuncheva and C. J. Whitaker, "Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy," *Mach. Learn.*, vol. 51, no. 2, pp. 181–207, 2003.
[20] A. Temko, A. Sarkar, and G. Lightbody, "Detection of seizures in intracranial EEG: UPenn and Mayo Clinic's seizure detection challenge," in *Proc. IEEE EMBC*, Aug. 2015, pp. 6582–6585.
[21] G. B. Boylan, L. Burgoyne, C. Moore, B. O'Flaherty, J. M. Rennie, "An international survey of EEG use in the neonatal intensive care unit," *Acta Paediatrica*, vol. 99, no. 8, pp. 1150–1155, 2010.
[22] E. M. Thomas, A. Temko, W. P. Marnane, G. B. Boylan, and G. Lightbody, "Discriminative and generative classification techniques applied to automated neonatal seizure detection," *IEEE J. Biomed. Health Informat.*, vol. 17, no. 2, pp. 297–304, Mar. 2013.
[23] V. Vapnik, *Estimation of Dependences Based on Empirical Data*. New York, NY, USA: Springer-Verlag, 1982.
[24] I. Osorio *et al.*, "Performance reassessment of a real-time seizure-detection algorithm on long ECoG series," *Epilepsia*, vol. 43, no. 12, pp. 1522–1535, 2002.
[25] G. Cauwenberghs and T. Poggio, "Incremental and decremental support vector machine learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 13. 2001, pp. 409–415.
[26] A. Bordes, S. Ertekin, J. Weston, and L. Bottou, "Fast kernel classifiers with online and active learning," *J. Mach. Learn. Res.*, vol. 6, pp. 1579–1619, Sep. 2005.
[27] M. J. F. Gales, "Cluster adaptive training of hidden Markov models," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 4, pp. 417–428, Jul. 2000.
[28] X. Huang and K. F. Lee, "On speaker-independent, speaker-dependent, and speaker-adaptive speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 1, no. 2, pp. 150–157, Apr. 1993.
[29] F. J. Provost, T. Fawcett, and R. Kohavi, "The case against accuracy estimation for comparing induction algorithms," in *Proc. ICML*, 1998, pp. 445–453.
[30] S. J. Mason and N. E. Graham, "Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation," *Quart. J. Roy. Meteorol. Soc.*, vol. 128, pp. 2145–2166, Jul. 2002.
[31] A. H. Ansaria *et al.*, "Improved multi-stage neonatal seizure detection using a heuristic classifier and a data-driven post-processor," *Clin. Neurophysiol.*, vol. 127, no. 9, pp. 3014–3024, 2016.

[32] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digit. Signal Process.*, vol. 10, nos. 1–3, pp. 19–41, 2000.
[33] D. Reynolds, "Gaussian mixture models," in *Encyclopedia of Biometrics*. New York, NY, USA: Springer, 2009, pp. 659–663.
[34] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Stat. Soc.*, vol. 39, no. 1, pp. 1–38, 1977.
[35] J. Platt, "Probabilistic outputs for SVM and comparison to regularized likelihood methods," in *Advances in Large-Margin Classifiers*. Cambridge, MA, USA: MIT Press, 1999, pp. 61–74.
[36] J. A. Hanley and B. J. McNeil, "A method of comparing the areas under receiver operating characteristic curves derived from the same cases," *Radiology*, vol. 148, no. 3, pp. 839–843, 1983.

ANDRIY TEMKO (M'05–SM'12) received the B.E. degree in computer science from Dnipropetrovsk National University, Dnipro, Ukraine, in 2002, and the Ph.D. degree in telecommunication from the Universitat Politècnica de Catalunya, Barcelona, Spain, in 2008. Since 2008, he has been with the Irish Centre for Fetal and Neonatal Translational Research, University College Cork, Ireland. He developed and patented a novel cutting-edge neonatal seizure detection system, which has completed European multi-centre clinical trial toward its regulatory approval and clinical adoption. In 2017, he was named a Winner of the Kaggle Challenge, and received 10,000 for the First Prize, for his work in predicting seizures in the human brain through long-term EEG recordings, organized by National Institute of Health, American Epilepsy Society, and Melbourne University. His research interests include acoustic and physiological signal processing, clinical decision support tools, and applications of machine learning for signal processing.

ACHINTYA KR. SARKAR received the master's degree in instrumentation and control engineering from Punjab Technical University, India, in 2006, and the Ph.D. degree from IIT Madras, Madras, India, in 2011. From 2011 to 2015, he was a Researcher in several laboratories, such as the Laboratoire Informatique d'Avignon, France, the LIMSI-CNRS, France, and the Department of Electrical and Electronic Engineering, University College Cork, Ireland. He is currently a Post-Doctoral Researcher with the Department of Electronic Systems, Aalborg University, Denmark. His research interests include speaker recognition, spoofing countermeasure, and seizure detection.

GERALDINE B. BOYLAN received the M.Sc. degree in physiology and the Ph.D. degree in clinical medicine from the University College London, London, U.K. She was a Clinical Scientist in neonatal medicine with the Kings College Hospital, London, from 1996 to 2001. She is currently a Professor of paediatrics with the Department of Paediatrics and Child Health, University College Cork, Cork, Ireland. Her research interest includes on accurately diagnosing seizures or fits in newborn babies by monitoring electrical brain activity and studies of blood flow regulation during neonatal seizures.

SEAN MATHIESON received the B.Sc. degree in biology from the University of Bath and the M.Sc. degree in neuroscience from UCL. He trained and was a Clinical Physiologist with the Neurophysiology Department, Great Ormond Street Hospital, from 2001 to 2009. Then, he was with UCL/UCLH, London, until 2016, on two research grants to test the ANSeR automated seizure detection algorithm for neonatal EEG, developed by researchers at UCC, Cork, Ireland. During this time, he studied the Ph.D. degree, which involved analysis of various aspects of the performance of the ANSeR algorithm. He is currently with UCC on the ENRICH study to investigate the effect of environmental enrichment on infant sleep and cognitive development.

WILLIAM P. MARNANE received the B.E. degree in electrical engineering from the National University of Ireland, Cork, Ireland, in 1984, and the Ph.D. degree from the University of Oxford, Oxford, U.K., in 1989. He was a Visiting Researcher with the Electronic Devices Research Group, Department of Physics, University of Linkping. He is currently a Professor with the Department of Electrical and Electronic Engineering, University College Cork, Cork, Ireland. His research interests include biomedical signal processing and digital design for DSP, coding, and cryptography.

GORDON LIGHTBODY received the M.Eng. (Hons.) and Ph.D. degrees in electrical and electronic engineering from Queen's University Belfast in 1989 and 1993, respectively. He is currently a Senior-Lecturer in control engineering with University College Cork, Ireland, and a founding Principal Investigator of the Science Foundation Ireland in both the Infant and the Marei research centres. His current research interests include artificial intelligence techniques for intelligent control and signal-processing, focusing on energy/power and biomedical applications. He is currently an Associate Editor of the Elsevier Journal, *Control Engineering Practice*.