ELSEVIER

Neurocomputing 30 (2000) 201–218

NEUROCOMPUTING

www.elsevier.com/locate/neucom

# Recurrent neural network based prediction of epileptic seizures in intra- and extracranial EEG

Arthur Petrosian[a,*], Danil Prokhorov[b], Richard Homan[a], Richard Dasheiff[a], Donald Wunsch II[c]

[a]*Health Sciences Center, Department of Neurology, Texas Tech University, 3601 4th Street, Lubbock, TX 79430, USA*

[b]*Ford Research Laboratory, Dearborn, MI, USA*

[c]*Applied Computational Intelligence Laboratory, Department of Electrical Engineering, Texas Tech University, Lubbock, TX 79430, USA*

Received 5 February 1998; accepted 22 March 1999

**Abstract**

Predicting the onset of epileptic seizure is an important and difficult biomedical problem, which has attracted substantial attention of the intelligent computing community over the past two decades. We apply recurrent neural networks (RNN) combined with signal wavelet decomposition to the problem. We input raw EEG and its wavelet-decomposed subbands into RNN training/testing, as opposed to specific signal features extracted from EEG. To the best of our knowledge this approach has never been attempted before. The data used included both scalp and intracranial EEG recordings obtained from two epileptic patients. We demonstrate that the existence of a "preictal" stage (immediately preceding seizure) of some minutes duration is quite feasible. 

*Keywords:* EEG; Epileptic seizure; Recurrent neural network; Wavelet transform

## 1. Introduction

Electroencephalography provides key information for the interictal definition of epilepsies and localizing *ictal* (epileptic seizure) onset [5,8]. The ability to reliably define a "preictal" state, on the other hand, would allow investigation of underlying

* Corresponding author. Tel.: + 1-806-743-2495; fax: + 1-806-743-1668.
*E-mail address:* neuaap@ttuhsc.edu (A. Petrosian)

0925-2312/00/$ - see front matter 
PII: S0925-2312(99)00126-5

neurochemical, metabolic and other changes through application of functional neuroimaging techniques. Use of those techniques such as MR spectroscopy, or PET and SPECT with radioligands injected during the *preictal* state, could define neurotransmitter or neurochemical changes leading to seizures. In turn, this could lead to improved anti-epileptic drug management, or, alternatively, early warning and electrical interruption systems [23,33].

Over the past two decades much research has been done with the use of conventional temporal and frequency analyses measures in the detection of epileptiform activity in EEGs and comparatively good results have been obtained [4,11,12,21,31]. Many investigators, at the same time, have focused their studies on quantitative characterization of underlying nonlinear dynamical systems (*chaos*) based on some evidence of a deterministic nature of the EEG dynamics [17]. The complexity measures of underlying EEG dynamics, such as *correlation dimension, Kolmogorov entropy*, and *Lyapunov exponents*, have been derived and investigated [1,6,7,14,19,16]. It has been reported, in particular, that the value of the correlation dimension computed along the temporal axis drops during an epileptic seizure. Yet the complexity of spatial dynamics was quantified by the "global" correlation exponent obtained from multichannel EEG [6]. It should be emphasized however, that the application of related mathematical algorithms are restricted by (a) the requirements of long-term (theoretically infinite) stationary time series considerations, and (b) an extremely high sensitivity to the noise, both electrical and physiologic in its nature. Therefore, it is necessary to systematically test data for stationarity before applying tests for chaotic behavior. This may often not be the case, particularly in determining preictal states from epileptic EEG. Our efforts were therefore directed to constructing measures, which could reflect short-term signal "textural" complexity conditions which may then result in seizures [24–26]. In [26] we showed the feasibility of using *local texture features* in conjunction with a *wavelet transform* for predicting the epileptic seizure well prior to its onset. To confirm those results we correlated them with the behavior of signal basic algorithmic and informational complexity measures, such as *Kolmogorov complexity*, and *fractal dimension* [24]. Note that these studies were performed on *intracranial* EEGs – the seizure recordings obtained with depth electrodes, surgically implanted into the patient's brain. It is known that these cortical EEGs differ from the ones recorded from the surface of the scalp (*extracranial*) in that they contain more power in higher frequency range. This is mainly due to the fact that high frequencies tend to be spatially coherent over small cortical surface areas [22]. The scull acts as a low-pass filter further attenuating fast frequencies.

In the present study we explore the ability of specifically designed and trained *recurrent neural networks* (RNN), combined with wavelet preprocessing, to predict the onset of epileptic seizures both on scalp and intracranial recordings. We will not attempt to extract specific signal features from EEG and will instead input the original EEGs and their wavelet filtered subbands by themselves into RNN for training/testing purposes. Although neural networks have been successfully applied in the past for epileptic spike and epileptic seizure detection problems [15,34,28], they have not been extensively investigated for seizure prediction purposes. In fact, we are unaware of any published results up to date with the use of neural networks trained on

raw EEG data for long-term seizure prediction. We attribute this to the challenging computational requirements of dealing with huge volumes of raw EEG data and the appropriate choice of related network training procedures. In this paper we use network training procedure based on the *decoupled extended Kalman filter* (DEKF) algorithm which has been shown to be superior to the standard gradient descent training algorithms in nonlinear dynamical systems control [30,29]. Previously, DEKF-trained RNNs have been successfully applied to a number of other complex problems, such as stock market trend predictions and automotive engine misfire detection [9,20,36,32]. Our data set consisted of four intracranial seizure recordings of one epileptic patient and six extracranial recordings obtained from another patient. The data extracted from selected channels of these multichannel EEGs contained up to 10 min of recordings immediately preceding the seizure onset. Our goal was to demonstrate the existence of several minute long preictal stages rather than to develop an on-line detection tool. We have been able to show that those preictal stages are in fact apparent when the DEKF methods of training RNNs were applied to wavelet decomposed high-pass subsignals of intracranial recordings. On surface electrode data the results were not as apparent. However, with a more extensive training procedure (described in detail below) we have been able to achieve acceptable results on all six scalp recordings using the *leave-one-out* testing method.

## 2. EEG data acquisition and wavelet decomposition

The EEG data used in this study were obtained from two patients who were undergoing long-term electrophysiological monitoring for epilepsy. The first set of intracranial EEG was obtained from a patient with depth electrodes surgically implanted into the epileptic focus. A 32-channel EEG with a sampling rate of 200 Hz was recorded using the *Stellate Monitoring System*. These data included four separate seizures which occurred on different days. Corresponding preictal single-channel recordings considered in this study are referred below as I1 (intracranial-1, used for training), and I2, I3, I4 (used for testing). Along with these we also used three other segments of intracranial "normal" EEG (N1, N2, and N3) that contained no interictal spikes or other abnormal activities. Each of these segments contained about 2 min of data, with N1 being used for training and N2, N3 for testing. As an example, two subsegments of EEG data from I1 and N1 are presented in Fig. 1. The corresponding wavelet decomposed low-pass and high-pass subsignals are denoted below with $*$.lp and $*$.hp extensions.

On one particular recording (that contained I1) the visual-based segmentation into interictal, preictal-1 (seizure non-inevitable), preictal-2 (seizure inevitable), ictal, and postictal stages was performed by an expert epileptologist. The epileptologist's determination of the various stages was based on the review of several days of EEG data which included multiple clinical seizures and numerous electrical events which did not proceed to clinically recognizable seizures. This entire recording was investigated in [25] using a novel *texture analysis* technique. Texture features relate to the presence of organized structures within EEG epochs and may reflect the complexity and nature of

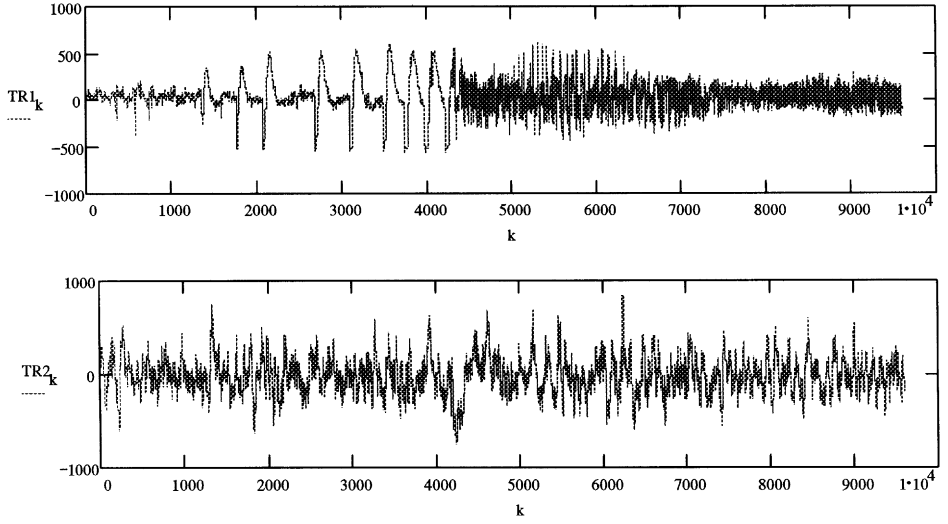



Fig. 1. Two 48 s epochs of intracranial EEG: at the top the transition from preictal to seizure occurs at around the 4400 sample mark (only preictal data were used for training); at the bottom is a segment of "normal" data.

transitions which occur in the EEG and which are not visible to a human expert. In this particular recording, noticeable temporal changes occurred in those features' behavior some 90 s prior to seizure onset (as we shall see below, precisely the same localization was achieved with RNN).

This technique was further enhanced using directionally decomposed low-pass and high-pass subbands obtained by wavelet transform [26]. Multiresolution wavelet representation provides a simple hierarchical framework for interpreting the signal information [18,2]. At different resolutions, the signal details may characterize different physiological processes in the signal source. The ability of wavelets to extract and localize those specific transient patterns makes them a natural complement to the application of RNN. In this study we apply wavelet transform to the signal to separate its frequency band into low-pass and high-pass subbands. In fact, we demonstrate that this decomposition was essential for achieving the network's steady behavior on our data. The choice of a particular optimal shape of the mother wavelet [3] for signal decomposition requires extensive investigation which was beyond the scope of this study. We used Daubechies kernel function [3] that has good localizing properties both in temporal and frequency domains.

The extracranial EEG was recorded using a *TECA 1121* (*Vickers Medical*) with scalp electrode placement according to the International 10–20 System [5]. The long-term monitoring for epilepsy was performed over 2 days during which the patient experienced six clinical seizures. A single bipolar longitudinal, parasagital and temporal chain montage was employed ("double banana"). Split screen video monitoring with both head and entire body views was included. Synchronization of video and EEG was performed with a digital readout on the video and a time scribe on the paper print out.

Visual analysis of these multichannel paper EEGs were performed by two expert epileptologists to localize seizure onsets and to mark segments containing various artifacts. In addition, a few "critical" channels that are most relevant to a particular ictogenesis were identified. Those signals along with their wavelet decomposed subsignals were used for RNN training and testing. Note that the corresponding digitized recordings contained longer preictal states than those in intracranial data. However, for consistency purposes we usually considered only about 4 min of single-channel preictal (preceding the seizures) data on each of those recordings. They are referred below as E1–E6 (extracranial 1–6).

## 3. Training recurrent neural networks

Recurrent neural networks (RNNs) considered in this paper belong to the well-known type of *discrete-time recurrent multilayer perceptrons* (see, e.g., [13]). Temporal representation capabilities of these networks can be significantly better than those of purely feedforward multilayer perceptrons or feedforward networks with tapped-delay lines. Unlike other networks, RNNs are capable of representing and encoding strongly hidden states, i.e. states in which a network's output depends on an arbitrary number of previous inputs. However, despite their advantageous architecture (Fig. 2) the RNNs have not been widely used in practical applications due to the lack of an efficient and universal training method. Recent developments in the use of parameter-based extended Kalman filter algorithm in RNNs training procedure have proven these architectures to be of great practical value [13,30,10,20,32]. We make use of this algorithm which is summarized below in Section 7. Note however, that unlike all the previous research we dealt in this work with a challenging and unique task of training RNN on a target which is not well defined. That is, we set the RNN output to remain steadily negative during interictal and to jump to and stay in steadily positive values with the transition to a preictal stage, without actually utilizing any preliminary

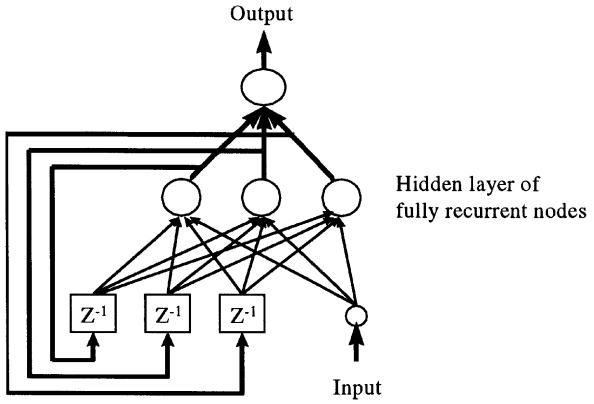



Fig. 2. Recurrent neural network architecture.

information as to precisely when and how this transition between interictal and preictal stages occurs. Namely, the desired network output (target) was set to be as close to $\pm 0.85$ as possible, $-0.85$ being a target value for interictal stage and $+0.85$ for preictal stage (we found the training process to be slightly faster when $\pm 0.85$ values were used instead of $\pm 1$). This choice of the desired output function, albeit simple and arbitrary, helped us in best network performances selection process using simple criteria. The latter included the quantitative parameters of the percentages of "false positives" and "false negatives" during interictal and preictal stages accordingly, as well as subjective visual analysis of the network output plots.

We have used the multi-stream EKF-based training of RNN on single-channel intra- and extracranial EEG recordings with a typical length ranging from 20,000 to 48,000 samples (the length was half as much on wavelet decomposed subsignals). We uniformly split those training sets into 50 nonoverlapping regions. These regions were used to allocate 50 streams (one stream per region). The length of each stream, or *training length*, henceforth referred to as the parameter $L$, was varied from 10 to 20 consecutive training pairs. Each training stream started from an entry point chosen randomly within a corresponding region of the training set. Note here, that before starting actual training of the RNN on $L$ training pairs, we performed *priming* of the network, i.e. the RNN was fed with a sequence of consecutive samples of length PL (*priming length*) in order to build up its internal states. This priming length was the same for each stream. It is normally proportional to the number of recurrent neurons in the RNN, and we put $\mathrm{PL} = 5, \ldots, 10$.

After 50xPL total samples were processed, the first weights' update was initiated. It was based on the first 50 training pairs spread over 50 different regions of the whole training set. After the first 50 pairs had been processed, we proceeded on to the next 50 pairs. We repeated such an updating of weights until the last 50 pairs in the streams were used. Then we chose new entry points and continued this procedure (including priming) until an acceptable performance was attained.

Our final network's architecture was a one-hidden-layer RNN with one or two inputs, 10 or 15 fully recurrent hidden neurons and one output neuron. Although we experimented extensively with larger networks with two hidden layers (the largest RNN used had the $1 \times 20\mathrm{r} \times 10\mathrm{r} \times 1$ architecture, where "r" stands for a recurrent neuron), we did not observe a noticeable improvement in training and testing performance.

Input to the network was normalized by subtracting a mean value and dividing by a standard deviation of all inputs in the training/testing set. Recurrent and output nodes had the common bipolar sigmoid nonlinearity. The set of matrices $H_i$ was obtained by the truncated backpropagation through time with the depth 20 (see Section 7), that is we did not use more than 20 copies of the network to accumulate appropriate derivatives in matrices $H_i$.

Training usually lasted several hundred passes, where one pass corresponded to a complete processing of all 50 streams. The programs were run on two 200 MHz *PentiumPro* workstations with a training procedure of 750 passes on a single raw EEG recording lasting for 2–3 h. The learning rate $\eta(k)$ and noise factor $q(k)$ were

interactively controlled through this procedure, and the corresponding best network parameters and weights were being saved for testing. The network's output was minimized by commonly used mean-square errors $e(k)^2$ on the whole training/testing data. This most important quantitative criterion along with two other parameters — the percentages of false alarms (positive outputs during interictal stage) and missed detections (negative outputs during preictal stage), were usually sufficient to assess the obtained network performance even before plotting its output (Figs. 3–9). Naturally, the best training/testing output was most frequently observed when all three criteria had low enough values.

The training was implemented on both raw EEG signal and wavelet decomposed subsignals (see Sections 4 and 5). The results revealed an apparent improvement of network behavior when applied to wavelet filtered subbands as opposed to the original data.

## 4. Results on intracranial data

### *4.1. Experiments on raw EEG recordings*

The above-described EEG data file I1 containing 34,000 samples – 170 s of recording immediately preceding the seizure onset – was used as the first training set. We specified targets for the network's training by splitting this file into two halves and assigning constant values of $\pm 0.85$ to the first and second segments accordingly (we followed this same interictal/preictal encoding rule throughout this study without regard to any specific qualitative signal information). Note, that on this particular recording we did attempt to experiment with the negative-to-positive transition marks in accordance with previously obtained transitional changes in signal texture and complexity characteristics [24–26]. While in some cases we observed certain improvements in RNN behavior, we found it insignificant and did not follow this approach.

The experiments with the I1 data as the only input for training resulted in the RNN reporting too many false positives (seizure onset prediction alarms) on "normal" EEG segments. We therefore trained the RNN on I1 along with another segment N1 of EEG data that contained no interictal spikes. We attached the 20,000 samples of the file N1 to augment the original training set I1, thus totaling 54,000 training samples. We then tested the trained network on the files containing other seizure recordings of the same patient. Typical results of the network's testing output on I2 is presented in Fig. 3. The results on I3 and I4 were similarly poor. It can be seen, in particular, that on I2 the RNN output jumps to the positive values even before the designated transition mark from $-0.85$ to $+0.85$, but then the network's output drops down to the negative area just prior to seizure onset. It should be noted, however, that this network remained steadily negative with only 5% (out of the total number of samples) of false outbursts when tested on N2 file (over 4 min of "normal" EEG segment).

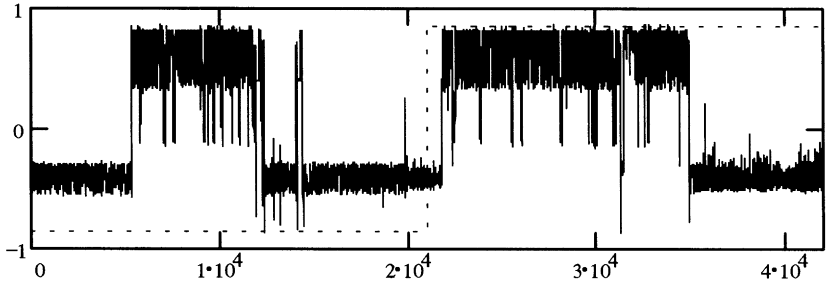



Fig. 3. Network testing results on original I2 EEG recording. The horizontal lines at $\pm 0.85$ indicate the desirable network output and the vertical line between them is a conditional transition boundary from interictal to preictal stage.

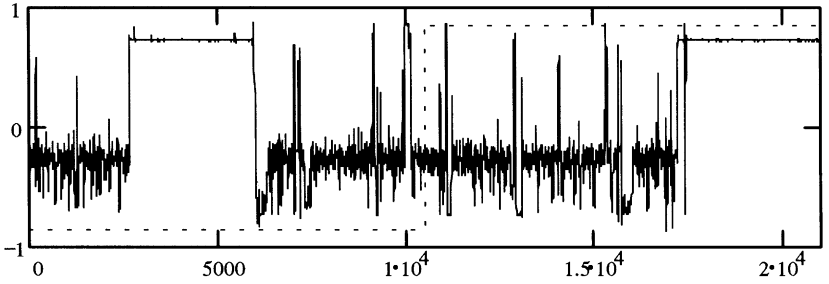



Fig. 4. The graph represents the results of network output during testing on I2.lp data.

### *4.2. Experiments on wavelet decomposed subsignals*

In the next phase of experiments we fed the RNN with the wavelet filtered data. We used the "daub4" filter [3] to decompose the original EEG data into its coherent component and noisy residue, i.e. the low- and high-pass subbands (referred below to as file extensions ∗.lp and ∗.hp, respectively). For example, the file I1 was decomposed into files I1.lp and I1.hp and similar to the above the training set was augmented by the files N1.lp and N1.hp.

We first trained the RNN on both I1.lp and N1.lp (27,000 samples in total). Results of testing on I2.lp and N2.lp are illustrated in Fig. 4. As compared to Fig. 3, the network's output contains some steady segment, distinctive of a preictal stage just prior to seizure onset. However, a similar positive output segment is seen during interictal stage too. Yet, the network exhibits somewhat "noisier" behavior outside of those segments. Nevertheless, the overall results of testing the RNN were better when they were trained on these coherent subbands rather than on original data.

We obtained the best prediction results when using high-pass subsignals (∗.hp files). Our training set included I1.hp and N1.hp, and the training performance is shown in

Fig. 5. The results of testing on I3.hp augmented by N3.hp are presented in Fig. 6 (top). It demonstrates that the network's output not only remained steadily negative within both interictal and "normal" stages (the file N3.hp was attached to the end of I3.hp, following its 15,000 samples), but is also steadily positive during the preictal stage. Stable transition of the network's output from negative to positive values occurs even some 15 s before the manually specified transition mark, which appears to be favorable. Testing results on I2.hp (shown in Fig. 6 (bottom)) and I4 were similar.

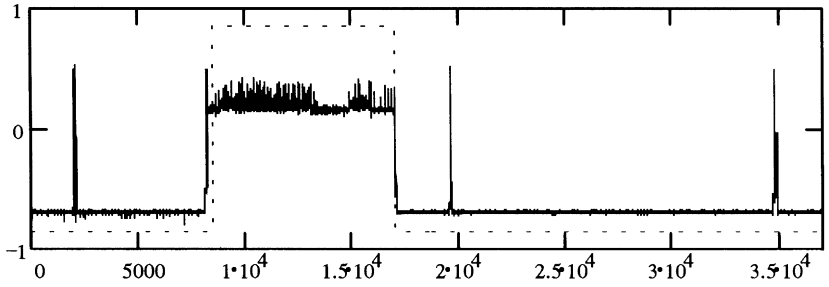



Fig. 5. Training on I1.hp (left part) and N1.hp (right part). It is apparent that the network was able to strictly follow training instruction for designated outputs of $-0.85$ and $+0.85$ (horizontal lines) except for one short-term outburst during interictal stage and two outbursts in "normal" stage.

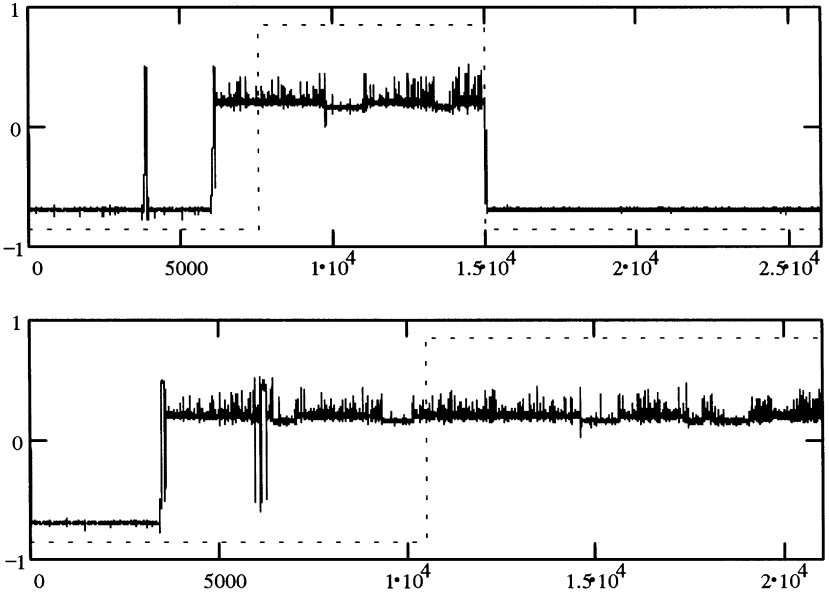



Fig. 6. Testing on I3.hp (top left part) and N3.hp (top right part) of the network trained as shown above in Fig. 5. At the bottom is testing on I2.hp. The favorable transition of the network's output from negative to positive values occurs much earlier than the designated transition boundary.

Again it is worthwhile mentioning an early negative-to-positive transition of network output values and their quite steady behavior.

## 5. Results on extracranial data

Our extracranial EEG data was comprised of six different seizure recordings with various lengths of preictal stages. However, for consistency purposes we cut those recordings to the segments that contained approximately 4 min of data preceding seizure onsets (E1–E6). In accordance with the results on intracranial data, the RNNs were trained to predict the seizure onset at about 2 min of the EEG recording.

We started experiments by training RNN on two original data files followed by testing on the remaining four. Despite some promising training outcomes, testing usually failed in those experiments. Then we used 3 + 3 strategy (training on three and testing on the other three) on the original, as well as on high-pass and low-pass subsignals. While the results on original data were again found to be unsatisfactory, we have been able to obtain some positive results on wavelet decomposed subsignals (see, e.g. Fig. 7). However, unlike the results described in Section 4, part B, we could not achieve apparent superior performance when using either set of wavelet decomposed subsignals. In some cases, RNN trained on high-pass subbands performed better during testing. In others, the low-pass subsignal-fed RNN were more reliable. Yet, we failed to reproduce those positive results with varying combinations of input files for training and testing. Therefore, it was decided to carry out leave-one-out test on all six subsignals (i.e., training on all files but one and testing on that remaining file). The leave-one-out method is a widely accepted testing procedure and is commonly used in experiments when the available data set is restricted and is insufficient for carrying out independent training and testing. We performed these tests both on low-pass (E1.lp–E6.lp) and high-pass (E1.hp–E6.hp) subsignals using different initial weights of the networks, varying training strategies with different number of neurons and hidden layers. Nevertheless, the RNN failed in two out of six tests on low-pass and four out of six tests on high-pass subsignals. Yet, we noticed that while training failed on E2.hp and E5.hp, the corresponding results on their low-pass counterparts, E2.lp and E5.lp, were successful (see Fig. 8). It was apparent that combining those decomposed subsignals as simultaneous inputs to the RNN could bring better results. Indeed, the conducted leave-one-out test using both low- and high-pass subbands as the network's inputs showed that it was possible to obtain such a RNN that would pass all six tests. Typical results of one of these tests performed are presented in Fig. 9. To perform a proper comparison, we used 15 neurons in the hidden layer of the RNN as opposed to 10-neuron RNNs used in two previous experiments. Thus the total number of weights in this 15-neuron RNN (61 weights) was identical to that of two 10-neuron RNNs combined ($2 \times 31$ weights).

After this improvement was achieved on combined inputs (E1.lp + E1.hp–E6.lp + E6.hp) we went back to raw EEG data segments (E1–E6) and carried out the

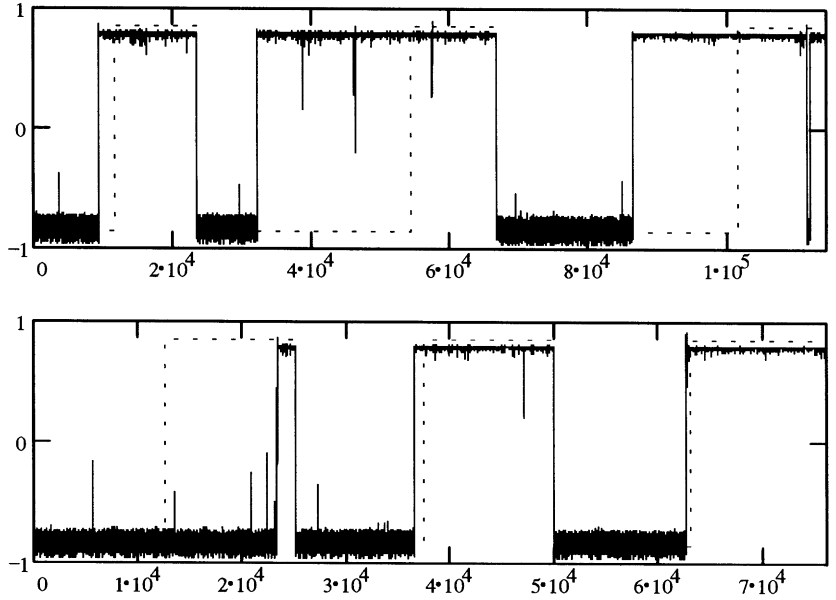



Fig. 7. Typical results of the experiments on the extracranial low-pass subbands. The upper plot shows the target and network's outputs on the low-pass filtered files that contain E1–E3.lp (from left to right) as their tails. The lower plot depicts testing performance of the same RNN on E4–E6.lp. As usual, the transition mark is set at the 2-min location prior to the seizure onset.

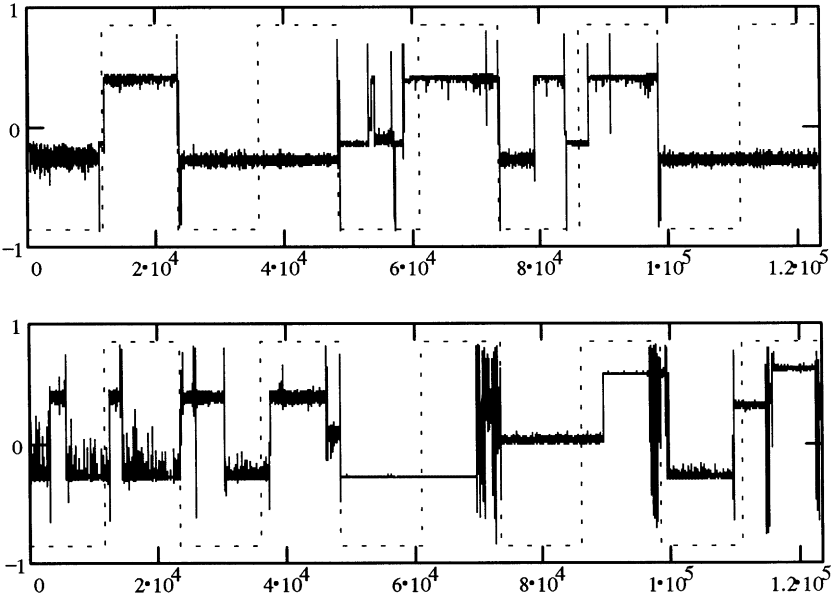



Fig. 8. Comparison of one of the leave-one-out tests with low-pass and high-pass subbands. The upper plot represents the results of training on E1–E5.lp, the lower one is during training on E1–E5.hp. Note that the network failed on E2.lp and E5.lp whereas it performed significantly better on their high-pass counterparts.

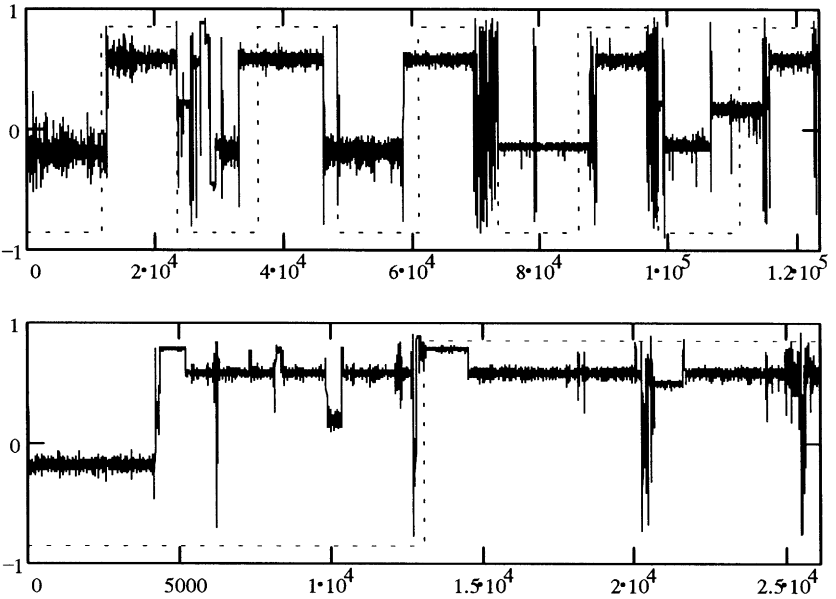



Fig. 9. Our typical results of one of the leave-one-out procedures for the RNN using the combined inputs. The upper graph is training results on E1–E5. Note oscillatory behavior just prior the seizure onsets on some of the recordings. This may be explained by existing artifacts in the original EEG. The lower graph is testing results of the same RNN on E6.

leave-one-out test procedure on this set. Three out of six tests failed, showing that both signal decomposition and combining obtained subsignals was essential in getting better network performance.

## 6. Discussion and conclusions

We have presented encouraging results in predicting the onset of epileptic seizures obtained with an advanced neural network approach and wavelet preprocessing. Consistent with previous results obtained using texture and complexity analyses, we have shown that the existence of a preictal stage of some minutes duration is quite feasible. As it was demonstrated on intracranial EEG recordings, it is the changes in the signal's high-frequency component ("noisy residue") that makes this prediction feasible. It is of interest to compare Figs. 3, 4, and 6, that show testing results on I2 recording. Obviously, the most stable network output behavior is the one shown in Fig. 6. Note also, that all these figures demonstrate some transitional change in the signal at a time instance of around 60 s (1200 samples of original signal) from the beginning of the recording. However, as it is seen in Fig. 6, despite a few down slips within a brief time period the network was able to maintain the favorable output during testing on I2.hp.

The network behavior on surface electrode data was less stable and long-term prediction somewhat problematic. However, this should not be surprising as one would expect much more meaningful high-frequency component in cortical rather than scalp EEG. We also observed that training on intracranial data usually resulted in local minima that were relatively deeper than those attained with training on extracranial data. In such minima all the performance criteria mentioned in Section 4, part A, reached low values simultaneously, thereby facilitating the choice of the best weights. Nevertheless, some promising results were achieved on extracranial data with the use of wavelet decomposition as well.

There are several points that we wish to emphasize here:

(a) The transition from interictal to preictal stage was specified in this study to be abrupt, as it was designated by the network's outputs. While it is possible to justify such a behavior hypothesizing the existence of different attractors of underlying nonlinear dynamical system, there is no substantial evidence to date supporting that hypothesis [16]. On the other hand, the use of a step function as the network's target function allowed us to judge the network's performance by simple rules. Specifically, we watched for a relatively stable output of the network both in interictal and preictal segments and an abrupt transition from negative to positive values in between those segments. Nevertheless, applying alternative types of the network's target function, such as gradual linear (or nonlinear) increase in network output values could be potentially useful.

(b) Associated with (a) is the problem of specifying the location of a transitional interictal–preictal boundary to be used for network training (we mostly used the middle mark of EEG segments considered). Although we did try to use a more substantiated approach of choosing the boundary mark based on the behavior of various texture and complexity characteristics [24–26], we were unable to demonstrate improved network behavior. Whether this reflects independence of the network's operation from the previously mentioned characteristics remains to be clarified in future studies.

(c) As it was described in Section 3, the application of RNNs involves multiple parameters and degrees of freedom (training parameters, network's architecture, its initial weights, etc.). All these may affect the training process and may require a lot of experiments to select their optimal set. In particular, we tried to use much bigger values for the training length parameter $L$ (up to 100) and for the priming length, PL (up to 10,000) hoping to extract some longer-term transitional changes in the signal. Apart from drastically slowing down the computational process, it did not produce any discernible improvement over the presented results. There is still much work to be done on establishing a relationship between those parameters and the optimal performance of the RNNs in predicting epileptic seizures.

(d) Training and testing of RNNs will be required on larger cross-patient data sets of original EEG recordings, as well as on combinations of subsignals at different levels obtained with various wavelet packet transforms [2]. It will also be of interest to run and compare the networks' results on intra- and extracranial channels recorded simultaneously in the same patient.

## 7. Summary of the EKF training algorithm

The extended Kalman filter-based (EKF-based) training, among many other methods proposed, stands out as an effective and powerful tool for the control of nonlinear dynamical systems [30]. In essence, an EKF-based training can be viewed as a parameter identification problem for a nonlinear dynamical system (RNN). It adapts the weights of the network in a pattern-by-pattern fashion accumulating important training information in approximate error covariance matrices and providing individually adjusted updates for each of the network's weights.

Let us assume that all weights of the RNN are assembled in a vector $W$ of length $M$, and that this vector is split in several disjoint groups. We denote $W_i$ as a weight vector of the $i$th group of weights. The index $i$ runs from 1 to $G$, where $G$ is the total number of groups. Thus, we assume that $W_1 \cup W_2 \cup \cdots \cup W_G = W$ and $\dim(W_1) + \dim(W_2) + \cdots + \dim(W_G) = M$.

It is required by the EKF method that we compute derivatives of the RNN's $N$ outputs (rather than more commonly used derivatives of output errors) with respect to the weights [30]. These derivatives are obtained through *backpropagation* through time or its truncated version [13,35,36]. We accumulate and store the derivatives in a set of matrices $H_i$ (one matrix per group of weights), where each $H_i$ is of dimension $\dim(W_i) \times N$.

The EKF-based network truing procedure can be described by the following equations:

$$\begin{aligned} A(k) &= \left(\eta^{-1}(k)I + \sum_{i=1}^{G} H_i^{\mathrm{T}}(k)P_i(k)H_i(k)\right)^{-1}, \\ K_i(k) &= P_i(k)H_i(k)A(k), \\ W_i(k+1) &= W_i(k) + K_i(k)e(k), \\ P_i(k+1) &= P_i(k) - K_i(k)H_i^{\mathrm{T}}(k)P_i(k) + Q_i(k), \end{aligned} \tag{1}$$

where $\eta(k)$ is a scalar learning rate, $K_i(k)$ is the Kalman gain matrix for the $i$th group of weights, $e(k) = d(k) - y(k)$ is the $N \times 1$ error vector [$d(k)$ and $y(k)$ are the desired and actual outputs respectively, and $e^{\mathrm{T}}(k)e(k)/2$ forms the mean-square error to be minimized over time]. Further, $P_i(k)$ is the approximate error covariance matrix, $\dim(P_i(k)) = \dim(W_i) \times \dim(W_i)$, which models correlations between each pair of weights within the $i$th group of weights, and $Q_i(k)$ is a positive diagonal matrix that helps to avoid numerical divergence of the procedure and prevents getting stuck in a poor local minimum [29].

Grouping of the weights can be done in a variety of ways. We employed the grouping-by-node method, i.e. weights belonging to the same neuron were grouped together. Thus $G$ was equal to the number of neurons, and we often ignored unnecessary correlations between weights that belong to different neurons. This resulted in a significant reduction of computational complexity and storage requirements as the dimensionality of error covariance matrices $P_i$ was made significantly smaller than $M^2$ (the dimensionality in case of $G = 1$).

The matrices $P_i(0)$ were initialized as diagonal matrices with large enough diagonal elements with values of around 1000. Over the course of training during our computer program implementation the user-specified values of $\eta(k)$ and $q(k)$ [same for all neurons diagonal components of $Q_i(k)$] were changed typically from 0.01 to 1.0 and from 0.01 to $10^{-4}$, respectively.

With large EEG data files, a batch update of weights is advantageous over a pattern-by-pattern update since the network learns to simultaneously minimize error on an entire batch of patterns taken from different regions of the data segment. To combine efficiency of the EKF-based training with a batch-like update, without violating a consistency between the weights and the approximate error covariance matrices, we made use of a *multi-stream training* approach. It was first proposed and tested in [10] and can be described briefly as follows.

We assume that there are $S$ data streams. It is then useful to consider $S$ copies of the same RNN (i.e. weights are the same for all the copies). Each copy is assigned to a separate stream. We then apply each copy of the RNN to a training vector drawn from the corresponding stream. We obtain $S$ $(N \times 1)$-stream-error-vectors and $S$ sets of matrices $H_i$. We then concatenate all corresponding $H_i$ to form one $H_i$ of dimension $\dim(W_i) \times (N \times S)$, where $N \times S$ is a total number of columns. Concatenating all the stream error vectors results in a single error vector e of dimension $(N \times S) \times 1$, where $N \times S$ is a total number of rows. The number of columns in each of the Kalman gain matrices $K_i$ is increased $S$ times. The global scaling matrix $A$ also grows $S$ times in size in both dimensions. The increase in dimensionality of the $A$ matrix should be of concern, as it must be inverted. This could become a potential bottleneck of the training procedure, particularly, when both $N$ and $S$ are big. However, we have not experienced problems while applying this multi-stream EKF-based procedure to a variety of tasks [20,32]. We have found the *singular value decomposition* [27] to be the best method for inverting the matrices $A(k)$ of dimensions as big as $100 \times 100$.

Note that the allocation of data streams can also be done in various ways. It is often impractical to use too short a stream since the total number of streams $S$ can make the $A$ matrix inversion too burdensome. On the other hand, having too few streams may not bring a significant improvement over the one-stream (basic) EKF-based training. The experiments show, that faster training of at least an order of magnitude (in terms of number of passes) and much better generalization can be expected with a multi-stream EKF training as opposed to a one-stream training. This difference grows further with larger training sets, reaching the point when one-stream training is completely inadequate for data sets comprised of tens of thousands of training vectors [20].

## Acknowledgements

The authors are grateful to the Texas Tech University Health Sciences Center, Texas Tech Center for Applied Research, Ford Research Laboratory, and National Science Foundation for their financial support. The authors also express their

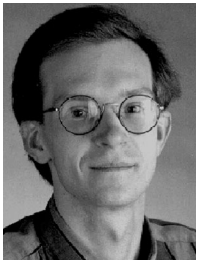
gratitude to the reviewers of this paper for their valuable comments, which have led to a significant improvement of the quality of this paper.

## References

[1] A. Babloyantz, Evidence for slow brain waves: a dynamical approach, Electroencephalogr. Clin. Neurophysiol. 78 (1991) 402–405.
[2] R. Coiffman, M. Wickerhauser, Wavelets and adapted waveform analysis, in: J. Benedetto, M. Frazier (Eds.), Wavelets, Mathematics and Applications, CRC Press, Boca Raton, 1994, pp. 399–423.
[3] I. Daubechies, The wavelet transform, time-frequency localization and signal analysis, IEEE Trans. Inform. Theory 36 (5) (1990) 961–1005.
[4] B.L.K. Davey et al., Expert system approach to detection of epileptiform activity in the EEG, Med. Biol. Eng. Comput. 27 (4) (1989) 365–370.
[5] F.H. Duffy, V.G. Iyer, W.W. Surwillo, Clinical Electroencephalography and Topographic Brain Mapping, Springer, New York, 1989, 304p.
[6] I. Dvorak, J. Klaschka, Modification of the Grassberger-Procaccia algorithm for estimating correlation exponent of chaotic systems with high embedding dimension, Phys. Lett. A 145 (1990) 225–231.
[7] J.-R. Eckmann, D. Ruelle, Recurrence plots of dynamical systems, Europhys. Lett. 4 (1987) 973–977.
[8] J.E. Engel, Seizures and Epilepsy, 2nd Edition, F.A. Davis Co., Philadelphia, PA, 1994.
[9] L. Feldkamp, D. Prokhorov, C. Eagen, F. Yuan, Enhanced multi-stream Kalman filter training for recurrent networks, in: J. Suykens, J. Vandewalle (Eds.), Nonlinear Modeling: Advanced Black-Box Techniques, Kluwer Academic Publishers, Dordrecht, 1998, pp. 29–53.
[10] L. Feldkamp, G. Puskorius, Training controllers for robustness: multi-stream DEKF, in: Proceedings of the World Congress on Computational Intelligence (WCCI), Orlando, FL, June/July, 1994, pp. 2377–2382.
[11] J. Gotman, Automatic recognition of epileptic seizure in the EEG, Electroencephalogr. Clin. Neurophysiol. 83 (1982) 271–280.
[12] J. Gotman, L. Wang, State-dependent spike detection: concepts and preliminary results, Electroencephalogr. Clin. Neurophysiol. 79 (1991) 11–19.
[13] D. Hush, B. Horne, Progress in supervised neural networks, IEEE Signal Process. Mag. 10 (1) (1993) 8–39.
[14] L.D. Iasemidis, J.C. Sackellares, H.P. Zaveri, W.J. Williams, Phase space topography of the electrocorticogram and the Lyapunov exponent in partial seizures, Brain Topogr. 2 (1990) 187–201.
[15] G. Jando et al., Pattern recognition of electroencephalogram by artificial neural networks, Electroencephalogr. Clin. Neurophysiol. 86 (1993) 100–109.
[16] B.H. Jansen, Nonlinear dynamics and quantitative EEG analysis, in: R.M. Dasheiff, D.J. Vincent (Eds.), Frontier Science in EEG: Continuous Waveform Analysis, EEG Suppl., Elsevier Science, 45, 1996, pp. 39–56.
[17] B. Kemp, F.H. Lopes Da Silva, Model-based analysis of neurophysiological signals, in: R. Weitkunat (Ed.), Digital Biosignal Processing, Elsevier Science Publishers B.V., Amsterdam, 1991, pp. 129–155.
[18] S.G. Mallat, A theory for multiresolution signal decomposition: the wavelet representation, IEEE Trans. Pattern Anal. Mach. Intell. 11 (7) (1989) 674–693.
[19] G. Mayer-Kress, Localized measures for nonstationary time-series of physiological data, Integrative Physiological Behav. Sci. 29 (3) (1994) 205–210.
[20] K. Marko, J. James, T. Feldkamp, G. Puskorius, L. Feldkamp, D. Prokhorov, Training recurrent networks for classification: realization of automotive engine diagnostics, in: Proceedings of the World Congress on Neural Networks (WCNN), San Diego, CA, September 1996, pp. 845–850.
[21] A.M. Murro et al., Computerized seizure detection of complex partial seizures, Electroencephalogr. Clin. Neurophysiol. 79 (1991) 330–333.
[22] P. Nunez, Electrical Fields of the Brain, Oxford U. Press, New York, 1981.
[23] E. Ott, C. Grebogi, J.A. Yorke, Controlling chaos, Phys. Rev. Lett. 64 (1990) 1196–1199.

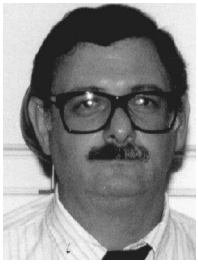
[24] A. Petrosian, Kolmogorov complexity of finite sequences and recognition of different preictal EEG patterns, The eigth IEEE Symposium on Computer-Based Medical Systems, Lubbock, TX, June 1995, pp. 212–217.
[25] A. Petrosian, R.W. Homan, The analysis of EEG texture content for seizure prediction, Proceedings of IEEE EMBS 16th Annual International Conference, Baltimore, Maryland, November 1994, pp. 231–232.
[26] A. Petrosian, R.W. Homan, S. Pemmaraju, S. Mitra, Wavelet-based texture analysis of EEG signal for prediction of epileptic seizure, SPIE Proceedings on Wavelet Applications in Signal and Image Processing, Vol. 2569, San-Diego, CA, July 1995, pp. 189–194.
[27] W.H. Press, S.A. Teukolsky, W.T. Vetterling, B.P. Flannery, Numerical Recipes in C: The Art of Scientific Computing, 2nd Edition, Cambridge University Press, New York, 1992.
[28] J.C. Principe, A. Zahalka, Noisy desired signal in transient detection using neural networks, Proceedings of the SPIE "Chaos in Biology and Medicine", San-Diego, CA, July 1993, pp. 292–300.
[29] G. Puskorius, L. Feldkamp, Decoupled extended Kalman filter training of feedforward layered networks, in: Proceedings of International Joint Conference Neural Networks (IJCNN), Seattle, WA, 1991, pp. I771–1777.
[30] G. Puskorius, L. Feldkamp, Neurocontrol of nonlinear dynamical systems with Kalman filter trained recurrent networks, IEEE Trans. Neural Networks 5 (2) (1994) 279–290.
[31] H. Qu, J. Gotman, Improvement in seizure detection performance by automatic adaptation to the EEG of each patient, Electroencephalogr. Clin. Neurophysiol. 86 (1993) 79–87.
[32] E. Saad, D. Prokhorov, D. Wunsch, Comparative study of stock trend prediction using time delay, recurrent and probabilistic neural networks, IEEE Trans. Neural Networks 9 (6) (1998) 1456–1470.
[33] S. Schiff, K. Jerger, D.H. Duong, T. Chang, M.L. Spano, W.L. Ditto, Controlling chaos in the brain, Nature 370 (1994) 615–620.
[34] W. Weng, K. Khorasani, An adaptive structure neural networks with application to EEG automatic seizure detection, Neural Networks 9 (7) (1996) 1223–1240.
[35] P.J. Werbos, Backpropagation through time: what it is and how to do it, Proc. IEEE 78 (10) (1990) 1550–1560.
[36] R. Williams, D. Zipser, Gradient-based learning algorithms for recurrent networks and their computational complexity, in: Chauvin, Rumelhart (Eds.), Backpropagation: Theory, Architectures, and Applications, Lawrence Erlbaum Associates, Hillsdale, NJ, 1995, pp. 433–485.

**Arthur Petrosian** received his M.S. degree in Mathematics from Moscow State University in 1983, and Ph.D. in Applied Mathematics from the Institute of Informatics & Automation of the National Armenian Academy of Sciences in 1989. He carried out his postdoctoral studies in mammographic image processing in the University of Michigan at Ann Arbor (1992–1993), and in EEG signal analysis in the Medical College of Ohio at Toledo (1993–1994). He joined the faculty at Texas Tech University Health Sciences Center as an Assistant Professor in 1994 and holds secondary appointments in the Department of Mathematics and Electrical Engineering, Texas Tech University. His research interests are in the areas of approximation and optimization theory and biomedical signal-image processing. He is a Senior Member of IEEE and a past member of The New York Academy of Sciences.

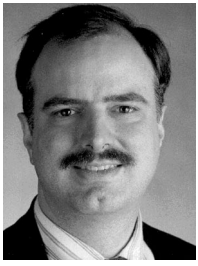


**Danil Prokhorov** received a Ph.D. in Electrical Engineering at Texas Tech University in 1997 and a Honors Diploma in Robotics from the State Academy of Aerospace Instrument Engineering (former LIAP), St. Petersburg, Russia, in 1992. He worked at the Institute for Informatics and Automation of Russian Academy of Sciences (former LIIAN), St. Petersburg, Russia, as a Research Engineer from 1992–1994. He is currently with Research Laboratory of Ford Motor Co., Dearborn, MI. His research interests are in adaptive critics, signal processing, system identification, control, and optimization based on various neural network approaches. He is a member of the International Neural Network Society.

**Richard Homan** received B.A. degree from Colgate University, Hamilton, New York, in 1962, and M.D. degree from State University of New York in 1966. He worked as an Assistant Professor of Neurology in UCLA Center for Health Sciences in 1970 and in Albert Einstein College of Medicine in 1972–1975. He also worked as an Assistant and Associate Professor in University of Texas Southwestern Medical School in Dallas in 1975–1989 and as a Chairman of Neurology Department in the Medical College of Ohio (1989–1994) and in Texas Tech University Health Sciences Center (1994–1997). His research interests include polypharmacy in seizure control and computerized methods for seizure prediction. He is a Vice-President of Texas Neuorologic Society and a member of American Academy of Neurology and Society for Neuroscience.



**Richard Dasheiff** received B.S. in Astronomy and Physics and M.D. in Medicine from the University of Maryland in 1972 and 1976. He completed his residency and postdoctoral fellowship in neurology in Duke University Medical Center in 1980 and 1982. Prior to joining Texas Tech in 1994 he worked as an Assistant Professor of Neurology in the University of Wisconsin-Madison and as an Associate Professor of Neurology and Psychiatry in the University of Pittsburgh Medical Center. He is currently a Professor and Director of Sleep and Epilepsy Centers at Texas Tech University Health Sciences Center. He is the President of the Epilepsy Foundation for the South Plains. He is a member of American Academy of Neurology, American Epilepsy Society, and American Medical Association.



**Donald Wunsch** received the Ph.D. in Electrical Engineering and the M.S. in Applied Mathematics from the University of Washington in 1991 and 1987, the B.S. in Applied Mathematics from the University of New Mexico in 1984, and completed a Humanities Honors Program at Seattle University in 1981. He is the Director of the Applied Computational Intelligence Laboratory at Texas Tech University. Prior to joining Texas Tech in 1993, he was Senior Principal Scientist at Boeing. He has also worked for International Laser Systems and Rockwell International. Current research activities include neural optimization, forecasting and control, fuzzy risk assessment for high-consequence surety, wind engineering, characterization of the cotton manufacturing process, intelligent agents, and Go. He is a member of the International Neural Network Society and a past member of the IEEE Neural Network Council.