A Comparison of Rule-Based and Machine Learning Methods for Classification of Spikes in EEG

Wolfgang Ganglberger, Gerhard Gritsch, Manfred M. Hartmann, Franz Fürbass, Hannes Perko, Ana Skupch, and Tilmann Kluge

Center for Health & Bioresources, AIT Austrian Institute of Technology GmbH, Donau-City-Straße 1, 1220 Vienna, Austria

Email: wolfgang.ganglberger@gmail.com; {gerhard.gritsch, manfred.hartmann, franz.fuerbass, hannes.perko, ana.skupch, tilmann.kluge}@ait.ac.at

Abstract—Diagnosis of epilepsy is based on the analysis of electroencephalogram (EEG) recordings. Essential epileptiform transients in the EEG are spikes, which are commonly marked manually by biomedical technical assistants which is very time-consuming and error-prone. Automatic spike detectors already exist but still have to be improved to better meet the needs of clinical experts. In this paper we discuss different automatic spike detection methods in order to improve the detection performance and to establish a user adjustable sensitivity parameter. The performances of a rule-based system, artificial neural networks (ANN) and random forests are investigated. For this retrospective study, data from an epilepsy-monitoring unit, including 12 patients comprising 130 hours recording time, were collected. The recordings were annotated by medical experts leading to a total of 5582 spikes. An artificial neural network exceeds the alternative methods in classifying the data set and achieves an average detection sensitivity of 44.1% and positive predictive value of 56.2% at a false detection rate of 19.8 per hour. Furthermore, the ANN also performs well in different sensitivity settings, enabling a user adjustable sensitivity parameter which helps the clinical experts to adjust the classifier to handle different application scenarios.

Index Terms—Epilepsy, spike detection, EEG, automatic, classification, rule-based, machine learning, minority class oversampling, artificial neural networks, random forests

I. Introduction

Approximately 1% of humans have epilepsy. In order to diagnose epilepsy, electroencephalogram (EEG) is recorded and analyzed. In clinical practice the analysis of EEG recordings is still often done manually, usually by biomedical technical assistants. Marking of abnormal EEG patterns as spikes and seizures is a fundamental part of the analysis. To increase review quality and time-efficiency, automatic tools able to assist the experts are essential.

In this paper we focus on automatic spike detection systems.

While there is great variability of what is considered a spike among electroencephalographers [1], Nochtar *et al*. defines a spike as a pointed transient, clearly distinguishable from EEG background, with a duration of 20-70ms and generally by having negative polarity relative to other scalp areas. Transients with similar characteristics but with durations up to 200ms are called sharp waves in keeping with this definition [2]. A fragment of EEG including an exemplary spike is shown in Fig. 1.

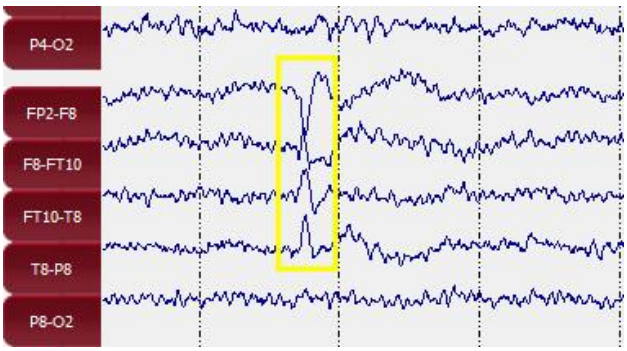

Fig. 1. Extract of an EEG. Intervals between two vertical, dashed lines correspond to one second. The yellow rectangle displays a pointed transient, clearly distinguishable from EEG background, called spike.

Various detection algorithms have been developed since the 1970s, basing on methods as measuring waveform morphology, detecting signal non-stationarity, power spectral analysis, wavelet analysis, neural networks or others. Many of published algorithms were reviewed in 2009 [3] but up to this day, a lack of a large, common standardized EEG dataset makes a performance comparison of different algorithms unfeasible.

At AIT we have developed an spike detection software encevis EpiSpike [4] that automatically marks spikes in EEG recordings.

Because the current system utilizes a rule-based classification method, the following improvements are intended to make. Firstly, we want to improve classification performance by reflecting the strength of human reviewers to jointly assess all spike features at once rather than deciding upon hard thresholds per feature. Secondly, a determination of a spike-quality measure for each transient is desired. In order to achieve these goals we have analyzed different machine learning approaches like Artificial Neural Nets (ANN) and Random Forests (RF). An increase of classification performance could raise acceptance of medical experts to

Manuscript received July 10, 2017; revised October 23, 2017.
doi:10.12720/jcm.12.10.589-595

use automatic tools. Also, these classifiers' outputs are able to serve as quality measures that enable the implementation of a user adjustable sensitivity parameter in order to adapt the sensitivity characteristic to handle different application scenarios or balance the high intra- and inter-patient variability of EEG signals and spike properties.

The paper is structured as follows. Section II begins with a description of the assembly of the data set and the annotation procedure. Then, the current EpiSpike detector and the machine learning algorithms are presented. In section III, detailed results and performance comparisons are demonstrated. A short discussion concludes the paper.

II. Methods

A. Data Set and Spike Annotation

EEG data used in this paper were retrospectively collected from recordings at the epilepsy-monitoring unit at the Second Neurological Department of the Neurological Center Rosenhuegel in Vienna. In total, 12 patients were selected and for all recordings, the international 10-20 electrode placement system and a sampling rate of 256 Hz were used.

Every EEG recording was manually examined by qualified and experienced EEG technologists. The examination comprised the annotation of spikes by marking the spike's peak in time, as well as the categorization of the annotated spikes into two quality-levels: 0.5 and 1. The technologists solely used their knowledge, experience and intuition for the annotation and the categorization-no formal criteria was defined beforehand. For reviewing the EEG signals, they used encevis 1.5 software. Furthermore, the technologists were hourly paid and no time limit for the examination was set.

For each patient, we observed EEG recording time. Summary statistics about the number of annotated spikes and their quality-level categorization were computed.

B. Data Processing and Feature Extraction

The following is a description of the processing of the EEG data and feature extraction implemented in encevis EpiSpike, with Fig. 2 serving as a graphical representation of this. At first, PureEEG algorithm [5], based on a neurophysiological model applying an iterative Bayesian estimation scheme, removes common scalp EEG artifacts like muscle, movement, line noise and loose electrode artifacts. With the resulting artifact-reduced EEG signal, bipolar longitudinal and transverse montages are created according to ACNS's proposal for clinical EEG montages [6]. For each bipolar channel, the signal is split into segments representing any kind of discharge with durations between 0.04 and 1.5 seconds and amplitudes exceeding 20 microvolts. The single-channel segments are then combined over several channels. For all these multi-channel segments, numeric features carrying information about the wave, e.g. the potential distribution, local context and signal morphology, are calculated. For the referential channel which exhibits the minimum of potential field, the wave-morphology is expressed by its amplitude, duration and slope. Features about field potential provide information about the spatial distribution and magnitude of the electrical potential discharges over electrodes. Local context features describe circumstances in which the wave occurred, e.g. during periods of high EEG, rhythmic activities or artifacts.

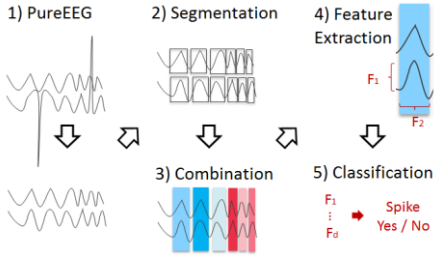

Fig. 2. Data processing and feature extraction. 1) artifact removal by PureEEG algorithm, 2) segmentation of signal into single-channel segments, 3) combination of single to multi-channel segments, 4) extraction of features for every multi-channel segment, 5) classification of multi-channel segments as spikes or non-spikes.

These multi-channel segment features are the basis for a classification into spikes and non-spikes.

C. Classification with EpiSpike Algorithm

A rule-based classification algorithm called EpiSpike integrated in encevis 1.5 software, was tested on the data set introduced above. In the development of the algorithm, the data set had been used also. In this algorithm, a set of conditions classifying the wave-segment based on their features is applied. Whenever a wave-segment meets these conditions, it gets classified as a spike by the algorithm.

D. Machine Learning Algorithms

As alternative methods for performing the spike classification task we used artificial neural networks and random forests, algorithms commonly used in the field of supervised machine learning. The inherent output for a binary classification of these methods is a value between 0 and 1, indicating how likely the instance being classified is affiliated with a certain class. This property is appealing for the spike classification task because these classification designs do not simply decide if a wave-segment is a spike or not and provide a quality measure of how likely the candidate is a spike.

1) Development data

In supervised machine learning, a learning algorithm derives a classification function from labeled data [7]. As development data, we used 24 $(= d)$ features of each wave-segment, described in section II D. To limit the amount of inessential data, we discarded all wave-segments that were classified as non-spikes in a rule-based classification with very sensitive settings. We obtained the development data set $X_{\text{develop, original}} \in \mathbb{R}^{m\,x\,24}$ with m as the number of remaining wave-segments. The development dataset was standardized so

that each column $C \in \mathbb{R}^{m\,x\,1}$ of $X_{\text{develop, original}}$ had mean 0 and variance 1.

Furthermore, in supervised learning, for each feature vector, a target value is required also. Because of different learning algorithms, two target values $y_{\text{expert, class}}$ and $y_{\text{expert,reg}}$ were generated for each wave-segment in the following way. If a spike was annotated by the expert between starting point and end point of a wave-segment, $y_{\text{expert, class}}$ was set to 1 and $y_{\text{expert,reg}}$ was set to the spike's quality-level, i.e. 0.5 or 1. If no spike was annotated in the time frame of a segment, both $y_{\text{expert, class}}$ and $y_{\text{expert,reg}}$ were set to 0.

Datasets with substantial divergence of the prior probabilities of different classes (i.e. non-spikes data are more prevalent than spike data) can significantly compromise the performance of standard learning algorithms [8]. Moreover, assuming each patient's EEG recording is different from the others, we expected the learning algorithms to better fit the patients with more data than those with fewer data. In order to acquire a more balanced dataset, we used minority class oversampling procedures in a leave-one-subject-out manner: For all patients P_i, we generated datasets D_i by using data of patients $P_j \neq P_i$ with $i, j \in \{1, \dots 12\}$ for oversampling. Data of patients with more than 500 annotated spikes were not used in oversampling.

Table II presents summary statistics about the over-sampled development datasets.

Oversampling methods used were a heuristic replication of data samples (*Duplication*) on the one hand and generation of synthetic samples (*Smote* and *Adasyn*) on the other hand.

For Duplication, we selected parts of the dataset, i.e. features and target values from spike data or from patients with few data, and copied it.

For Smote, the minority class is over-sampled by selecting minority class samples and introducing synthetic examples along the line segments joining the 5 minority class nearest neighbors [9].

For Adasyn, a weighted distribution for distinct minority class samples according to their level of difficulty in learning is used. For minority class samples that are harder to learn, more synthetic data is generated, which adaptively shifts the classification decision boundary toward the difficult example [10].

2) Artificial neural network

Two-layered feed-forward artificial neural networks with a single output neuron were used for ANN model development [11], see Fig. 3. The output $\tilde{y}_{NN}$ can be represented with the following equation:

$$\tilde{y}_{\text{NN}} = \tilde{g}\left(\sum_{j=1}^{h} w_{1j}^{(2)} \cdot g\left(\sum_{i=1}^{d} w_{ji}^{(1)} \cdot x_i + w_{j0}^{(1)}\right) + w_{11}^{(2)}\right)$$

where: d = number of features
h = number of hidden units
$w^{(i)}$ = weight matrix in layer $i \in \{1,2\}$
g = activation function for layer 1
$\tilde{g}$ = activation function for layer 2.

Two types of this net, a classification design and a regression design, with outputs called $\tilde{y}_{\text{NN, class}}$ and $\tilde{y}_{\text{NN, reg}}$ respectively, were applied. For both types, we set d = 24, h = 700, g as the hyperbolic tan-sigmoid function

$$g(n) = \frac{2}{1 + \exp(-2n)} - 1$$

and scaled conjugate gradient backpropagation [12] as the network training function.

For the classification design, $\tilde{g}$ was set to the logarithmic sigmoid function

$$\tilde{g}(n) = \frac{1}{1 + \exp(-n)}$$

and the cross-entropy function

$$p_c = \frac{1}{m}\sum_{k=1}^{m}[\, y_{\text{expert,class}}{}^{(k)} \cdot \ln\left(\tilde{y}_{\text{NN,class}}{}^{(k)}\right) + \\ + (1 - y_{\text{expert,class}}{}^{(k)} \cdot \ln\left(1 - \tilde{y}_{\text{NN,class}}{}^{(k)}\right)]$$

was utilized as performance function for training the network.

For the regression design, a linear activation function $\tilde{g}(n) = n$ was used for the output neuron and a performance function

$$p_c = \frac{1}{m}\sum_{k=1}^{m}(\, y_{\text{expert,reg}}{}^{(k)} - \tilde{y}_{\text{NN,reg}}{}^{(k)}\,)^2\,,$$

based on the mean squared error, was employed. For regression design, we did not use Smote and Adasyn oversampling.

In order to get a binary classification result, a threshold $y_{\text{th}} \in [0, 1]$ is applied to the output value $\tilde{y}_{\text{NN}}$. We define $y_{\text{NN}} = 1$ if $\tilde{y}_{\text{NN}} \geq y_{\text{th}}$ as spike and $y_{\text{NN}} = 0$ otherwise. Setting a threshold $y_{\text{th}} \in [0, 1]$, we define $y_{\text{NN}} = 1\ if\ \tilde{y}_{\text{NN}} \geq y_{\text{th}}$ and $y_{\text{NN}} = 0$ otherwise.

3) Random forest

Random forests are ensemble methods using tree-structured classifiers $\{h(x, \theta_k), k = 1, \dots\}$ where x is an input pattern and $\{\theta_k\}$ are independent identically distributed random vectors [13]. For training the classification model, RF algorithm builds multiple CART-like trees [14], each trained on a bootstrapped sample of the original training data. The number of trees was set to 50 in this study.

To grow each tree, 9 randomly selected features were used at each node. There have to be at least 10 observations for each tree leaf and misclassification costs are set equal for both classes. In RF binary classification, each tree classifies an input sample and $\tilde{y}_{\text{RF}} = \frac{N_{\text{Class1}}}{N_{\text{Trees}}}$, with N_{Class1} as the number of trees voting for class 1 and N_{Trees} as the total number of trees of the random forest, resulting in an output $\tilde{y}_{\text{RF}} \in [0, 1]$. Random Forest classification can be controlled by setting a threshold $y_{\text{th}} \in [0, 1]$. An instance with output $\tilde{y}_{\text{RF}} \geq y_{\text{th}}$ is ultimately classified as $y_{\text{RF}} = 1$ and $y_{\text{RF}} = 0$ otherwise.

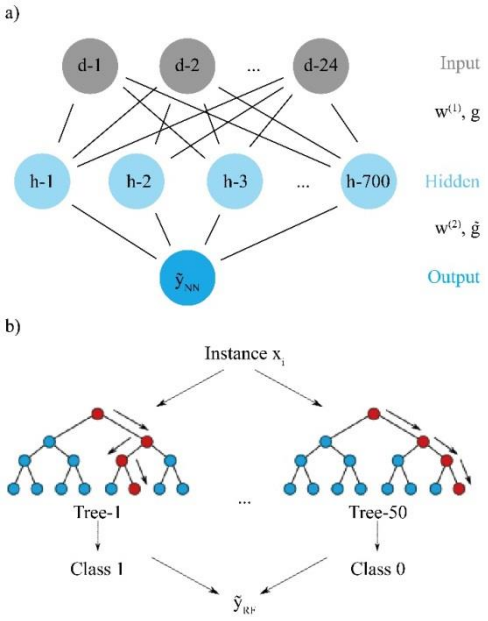

Fig. 3. a) Employed artificial neural network setting: 24 input neurons, 700 hidden units, one output neuron, weight matrices $w^{(1)}$and $w^{(2)}$ and activation functions g and $\tilde{g}$. b) Employed random forest setting: Each of 50 decision trees classifies an input instance x_i. Random forest output $\tilde{y}_{RF}$ is the proportion of number of trees voting for class 1 (= spikes).

E. Performance Metric and Evaluation

Metrics used for evaluating the performance of classification were:

True Positive (TP): $y_{\text{expert, class}} = 1$ and $y_{\text{classifier}} = 1$

False Positive (FP): $y_{\text{expert, class}} = 0$ and $y_{\text{classifier}} = 1$

False Negative (FN): $y_{\text{expert, class}} = 1$ and $y_{\text{classifier}} = 0$

Sensitivity (SE):

$$\text{SE} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

Positive Prediction Value (PPV):

$$\text{PPV} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

False Pos. Rate per hour (FP/h):

$$\text{FP/h} = \frac{\text{False Positives}}{\text{Recording time (h)}}$$

Since the inherent classifiers' outputs $\tilde{y}_{\text{NN,class}}$, $\tilde{y}_{\text{NN,reg}}$, $\tilde{y}_{\text{RF}} \in [0, 1]$ and the classification threshold y_{th} can be varied from 0 to 1, different paired values for sensitivity and PPV can be obtained. Those paired values can be represented as a curve, called PR-Curve (Precision-Recall) in the two-dimensional sensitivity and PPV space by plotting sensitivity on the y-axis and PPV on the x-axis. Interpolation between the points of the curve and trapezoidal integration let us obtain the area under the PR-Curve (PR-AUC) which is commonly used as a metric to define how an algorithm performs over the full sensitivity range [15].

Classification models were built and tested in a leave-one-subject-out manner: For training, original or oversampled (see section II D) data of 11 patients were used. All training data sets were used for the machine learning methods described above. The resulting models were tested with the original data of the left-out subject and PR-AUCs were calculated. Thus, all models were evaluated with a "new" patient's EEG data that was not used for model-development – representing real-world use cases.

The median PR-AUCs for the 12 patients were computed for every combination of training set - classification method.

The machine learning (ML) approach with highest median PR-AUCs was compared to the rule-based EpiSpike1.5 algorithm. EpiSpike was tested on each patient, resulting in performance pairs $SE_{\text{EpiSpike}}(P_i)$, $FP/h_{\text{EpiSpike}}(P_i)$ for each patient $P_i \in \{1,..12\}$ and the means over all patients $\overline{SE}_{\text{EpiSpike}}$ and $\overline{FP}_{\text{EpiSpike}}$ were computed. We selected an operating point of the ML model with $FP/h_{\text{ML}}(P_i) = FP/h_{\text{EpiSpike}}(P_i)$ and the corresponding $SE_{\text{ML}}(P_i)$ to compare the ML classification to EpiSpike. Thus, for each patient, a sensitivity change for fixed FP/h could be obtained.

For the ML algorithm, a mean performance curve over all patients was created by computing the mean sensitivity for all FP/h ∈ [0, 90]. For selected operating points on the curve, the sensitivity's empirical distribution for fixed FP/h over all patients was illustrated as a boxplot. Bottom and top of the boxes indicate the 25% and 75% quantiles, the whiskers represent the minimum and maximum of the data.

III. Results

A. Spike Data Set Summary Statistics

12 continuous EEG recordings of different patients were manually reviewed by the experts. The mean duration time of the recordings was 10.9 hours (min: 0.5h, max: 24h). In total, 5582 spikes were annotated by the expert, which resulted in a frequency of 43.05 marked spikes per hour. 2321 spikes were categorized as quality-level 0.5, 3261 as quality-level 1.

TABLE I. SPIKE DATA SET

Patient	Recording Time (h)	Annotated Spikes
1	0.5	173
2	3.9	155
3	0.6	146
4	2.8	121
5	2	61
6	0.53	148
7	0.74	173
8	24	118
9	24	377
10	24	1395
11	24	1032
12	24	1683

B. Development Data for Machine Learning

Table II shows the number of spike and non-spikes samples and their ratio for the datasets created for machine learning. With duplication, we increased the

number of spike samples by 1472 artificial spike samples. By utilizing Smote and Adasyn algorithms, 2944 artificial spikes were included to the data set.

TABLE II: DATA SETS USED FOR MACHINE LEARNING METHODS.

Data Set	# Spike Samples	# Non-Spike Samples	Non-Spike to Spike Ratio
$X_{develop,\ original}$	5582	264193	47.3 : 1
$X_{develop,\ duplicate}$	7054	264193	37.5 : 1
$X_{develop,\ smote}$	8526	264193	31 : 1
$X_{develop,\ adasyn}$	8526	264193	31 : 1

C. Performance

1) EpiSpike 1.5 – Performance

EpiSpike 1.5 classifier's mean patient performance was evaluated as described above. We obtained a mean sensitivity of 34.6% (with a standard deviation (SD) of 17.1%), mean PPV of 58.5% (SD =19.4%) and mean FP/h rate of 19.8 (SD = 24.4).

Sensitivity, exclusively evaluated with spikes of quality level 1, is 44.7% (SD = 15.0%).

2) Machine Learning Algorithms – Performances

Table III lists the mean area under the precision-recall curve (PR-AUC) for each combination of training set and classification method. The artificial neural net with classification design and $X_{develop,duplicate}$ used as training data (termed as ANN_α hereafter), with an AUC-PR of 0.425 tops the alternatives.

3) Performance Comparison

For all patients, performance values of EpiSpike and ANN_α with same False Positive Rates as EpiSpike are illustrated in Fig. 4. For the selected operating point of ANN_α (ANN_α *-OP)*, the sensitivity is greater than EpiSpike's in 8 out of 12 patients. For four patients, ANN_α 's sensitivity is more than twice as high as EpiSpike's while it's strongest decline is by a factor of 0.43 (from 63% to 35% for patient 8).

TABLE III: MEAN AREA UNDER PRECISION-RECALL CURVE FOR EACH COMBINATION OF TRAINING SET AND CLASSIFICATION METHOD.

	ANN Classification	ANN Regression	Random Forest
$X_{develop,\ original}$	0.399	0.380	0.388
$X_{develop,\ duplicate}$	0.425	0.404	0.390
$X_{develop,\ smote}$	0.391	-	0.390
$X_{develop,\ adasyn}$	0.389	-	0.392

ANN_α*-OP*'s mean sensitivity is 44.1% (SD = 12.9) and PPV is 56.2 % (SD = 30.2) which manifests an increase of 9.5 percentage points (PP) in SE and a decrease of PPV of 2.3 PP compared to EpiSpike 1.5. Standard deviation of ANN_α 's sensitivities is 13%, which corresponds to a reduction factor of 0.24 compared to EpiSpike's standard deviation of 17.1%.

Fig. 5 depicts the mean precision recall curve for ANN_α and the mean EpiSpike performance. Additionally, the empirical distributions of EpiSpike's FP rate and sensitivity and ANN_α's sensitivities for FP/h = 10, 30, 50 and 70 are presented as boxplots. The mean interquartile range for these four sensitivity-distributions is 29% - reflecting considerable different classification performances among the 12 patients. With a standard deviation of 1.7%, the magnitude of the interquartile ranges over the different FP/h values is relatively constant.

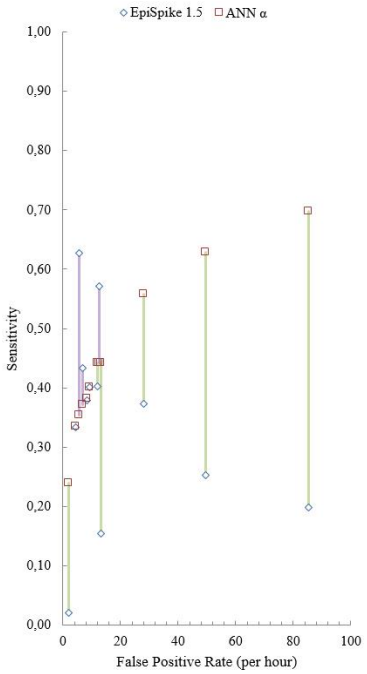

Fig. 4. Performance of spike detection of EpiSpike 1.5 and ANN_α for 12 patients. Performances of EpiSpike 1.5 for all patients were computed. For the same false positive rates, sensitivities of ANN_α were determined. Data point pairs from left to right correspond to the following patients: P6 – P5 – P8 – P2 – P12 – P9 – P1 – P11 – P4 – P10 – P3 – P7.

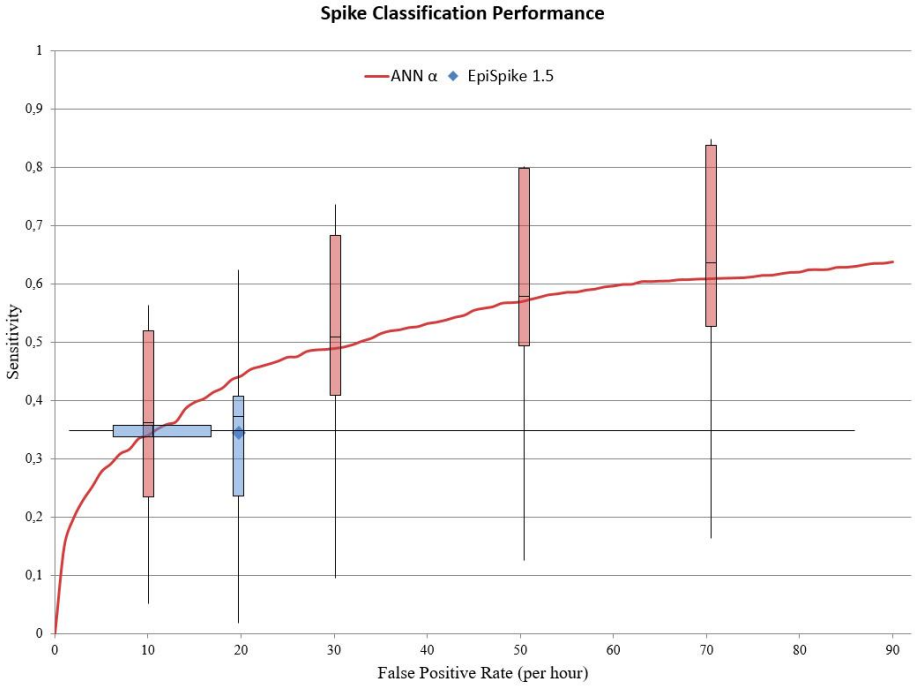

Fig. 5. Mean performance curve showing the relationship between sensitivity and FP rate. An operating point on the curve can be selected by choosing a corresponding classification threshold. For FP/h = 19.8, sensitivity is 44.1% and positive predictive value is 56.2% with artificial neural network.

The mean PR-curve's steep slope in the interval FP/h ∈ [0 10] indicates that sensitivity can be increased causing only few false positives (sensitivity 34% with 10 FP/h). A further increase of sensitivity leads to a relatively stronger increase in FP/h (e.g. sensitivity 60 % with 60 FP/h).

IV. Discussion

A gain in overall classification performance was achieved by using machine learning methods with artificial neural networks performing best on the dataset. The variance of performances for different patients was also lower using ANN, signifying a potential increase of stability for real world usages.

Addressing different prior class distributions by duplicating selected data samples led to models that performed better in classification than models generated with the original data set or with datasets with synthesized artificial samples.

With a sensitivity of 44.1% at a false positive rate of 19.8 per hour, the assessed performance is comparable to other spike detection systems and to inter-reader-agreements of experts [1]. We assume further improvements of spike detection systems to be possible, considering also progress in artificial intelligence methodologies [16].

High variances of the rule-based classification performances and the smooth ANN's precision-recall curve strengthen our opinion that a user adjustable sensitivity parameter could be a valuable feature in a future version of encevis. First implementation tests have shown favorable results and it is expected that clinical experts will be able to better handle different application scenarios with that feature.

Acknowledgment

We would like to thank Prof. Baumgartner form the Neurological Department Rosenhuegel at General Hospital Hietzing, Vienna, Austria for providing EEG data and detailed clinical findings.

References

[1] M. L. Scheuer, A. Bagic, and S. B. Wilson, "Spike detection: Inter-reader agreement and a statistical Turing test on a large data set," *Clin. Neurophysiol. Off. J. Int. Fed. Clin. Neurophysiol.*, vol. 128, no. 1, pp. 243–250, Jan. 2017.

[2] S. Noachtar, C. Binnie, J. Ebersole, F. Maugui ère, A. Sakamoto, and B. Westmoreland, "A glossary of terms most commonly used by clinical electroencephalographers and proposal for the report form for the EEG findings," *The International Federation of Clinical Neurophysiology, Electroencephalogr. Clin. Neurophysiol. Suppl.*, vol. 52, pp. 21–41, 1999.

[3] J. J. Halford, "Computerized epileptiform transient detection in the scalp electroencephalogram: Obstacles to progress and the example of computerized ECG interpretation," *Clin. Neurophysiol.*, vol. 120, no. 11, pp. 1909–1915, Nov. 2009.

[4] AIT Austrian Institute of Technology GmbH, [Online]. Available: http://www.encevis.com

[5] M. M. Hartmann, K. Schindler, T. A. Gebbink, G. Gritsch, and T. Kluge, "PureEEG: Automatic EEG artifact removal for epilepsy monitoring," *Neurophysiol. Clin. Neurophysiol.*, vol. 44, no. 5, pp. 479-490, Nov. 2014.

[6] American Clinical Neurophysiology Society, "Guideline 6: a proposal for standard montages to be used in clinical EEG," *Am. J. Electroneurodiagnostic Technol.*, vol. 46, no. 3, pp. 226–230, Sep. 2006.
[7] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of Machine Learning*, MIT Press, 2012.
[8] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009.
[9] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, no. 16, pp. 321–357, 2002.
[10] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in *Proc. IEEE Int. Jt. Conf. on World Congr. Comput. Intell.*, pp. 1322–1328, Jun. 2008.
[11] C. M. Bishop, Neural Networks for Pattern Recognition. Oxford university press, 1995 .
[12] M. F. Møller, "A scaled conjugate gradient algorithm for fast supervised learning," *Neural Netw.*, vol. 6.4, pp. 525–533, 1993.
[13] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
[14] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and Regression Trees*, CRC Press, 1984.
[15] Davis, Jesse, and Mark Goadrich., "The relationship between Precision-Recall and ROC curves," in *Proc. 23rd International Conference on Machine Learning*, 2006, pp. 233-240.
[16] S. H. Greenwald and C. K. Oertel, "Future directions in machine learning," *Front. Robot. AI*, no. 3, p. 79, Jan. 2017.

Wolfgang Ganglberger is currently completing a master of science program in biomedical engineering at TU Wien. He focuses on mathematical and computational biology. He earned a Bachelor of Science in mathematics in 2014 and finished a supplemental curriculum on innovation in 2017. At AIT, he has been developing algorithms for automatic spike detection in EEG.

Gerhard Gritsch received the MSc and the PhD from the Institute of Communications and Radio Frequency Engineering, TU Wien in 2001 and 2004, respectively. He has been developing algorithms for different applications in various engineering fields, e.g. wireless communication, laser scanning, traffic data acquisition and biomedical signal processing. Since 2010 he is with the brain signal analysis group of AIT focusing on EEG signal processing. His special research interests are spike detection and source localization. Gerhard Gritsch is author and co-author of more than 15 international papers.

Manfred Hartmann received the MSc from the institute of communications and radio frequency engineering, TU Wien in 2004. Since 2005 he is with the brain signal analysis group of AIT, where he is working on EEG signal processing algorithms with special focus on artifact reduction methods. Furthermore, he is involved in the software architectural design and the software QM processes. Manfred Hartmann is author and co-author of more than 20 international publications.

Franz Fürbass received the MSc from the Institute of Applied Information Processing and Communications, Graz University of Technology in 2006. Since 2009 he is with the brain signal analysis group of AIT, where he is working on EEG signal processing algorithms. His special research interest is automatic detection and classification of neurophysiological EEG patterns. At the moment he is finishing his PhD in the area of biomedical signal processing at the Institute of Electrodynamics, Microwave and Circuit Engineering, TU Wien. Franz Fürbass is author and co-author of more than 25 international publications.

Hannes Perko is Scientist at AIT. He got his diploma in electrical engineering from the TU Wien in 2002. During his studies he was focusing on multivariate signal analysis. Since 2002 Hannes Perko made major contributions to AITs EEG signal-processing algorithms. He is author and co-author of over 20 international publications.

Ana Skupch received an MSc from the institute of communications and radio frequency engineering in 2005 and an MSc from the institute of technical mathematics in 2007 (TU Wien). Since 2009 she is with the brain signal analysis group of AIT. Her special research interest is artifact detection and reduction. She is author and co-author of more than 15 international publications.

Tilmann Kluge is Senior Scientist at AIT and is leading the research field Bio-Signal Processing. He studied physics in Berlin and Boston and received his Ph.D. from the Max-Planck-Institute for Brain Research. He is author and co-author of more than 35 international publications.