On the overall ROC of multistage systems

Le Trung Thanh[†], Nguyen Thi Anh Dao[⋆,†], Nguyen Linh-Trung[†], and Ha Vu Le[†]

[†]University of Engineering and Technology, Vietnam National University Hanoi, Vietnam
[⋆]University of Technology and Logistics, Bac Ninh, Vietnam

***Abstract*—The receiver operating characteristic (ROC) curve is a useful tool to evaluate the performance of classifiers, and is widely used in signal detection, pattern recognition and machine learning. For complex object classification, multiple single classifiers are often used and they are concatenated into a multistage classification system. Thus, it is necessary to obtain the overal ROC curve, because the ROC curves of the individual classifiers are not useful for the overall system since it has multi-level decision thresholds. In this paper, a systematic approach was introduced for measuring the performance of multistage systems via estimating the *overall* ROC curve. Two new ROC models sharing the same properties of classical ROC curves were proposed, inspired by the Gaussian and logistic distributions. The models were then experimented on a recently introduced multistage system for epileptic spike classification from electroencephalogram data. Experimental results indicated that the proposed ROC models can be used for multistage classification systems.**

***Index Terms*—ROC, multistage system, multiple-decision threshold, epileptic spike, EEG.**

I. Introduction

Multistage systems composed of concatenated classifiers (a.k.a. detectors in signal processing) often provide better prediction, in comparison with single classifiers, probably because they accumulate the advantages of multiple algorithms. For example, in medical diagnosis, multistage systems are often used in detecting/classifying abnormal patterns in brain electrical activities [1], [2]. In natural language processing, multistage systems have also been successfully applied for recognition of emotional expression [3], [4]. Recently, in machine learning, the "trendy" deep learning models are fundamentally based on multistage architectures [5].

Multistage classification systems can be categorized into three main types: reject classifiers [6], cascade classifiers [7], and hierarchical classifiers [8]. In this paper, we are interested in the cascade classifiers. Among various methods to evaluate the performance of a classifier are the confusion matrix and the receiver operating characteristic (ROC) curve [9].

A confusion matrix may cover several common performance criteria, including accuracy, sensitivity, specificity, precision, and so forth. The ROC curve, which is a graphical plot presenting the range of probable predictions obtained by varying a decision threshold [10], is usually more efficient. One of the key advantages of the ROC curve over the confusion matrix is that it is insensitive to skew class distribution and misclassification costs. Thus, ROC-based analysis has become a useful tool to evaluate the performance of classifiers in various fields, including medical diagnosis [11].

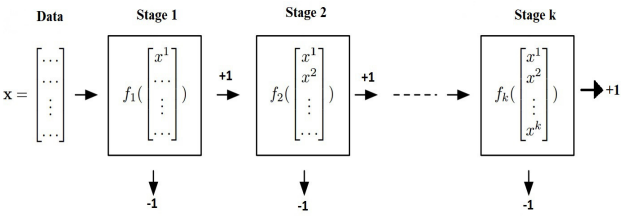

Fig. 1: A muti-stage classification system composed of k stages aimed at finding positive points (+1) out of the dataset.

Different approaches to estimating the ROC curve have been investigated, whether being non-parametric [12], [13], semi-parametric [14], [15], or parametric [16], [17]. A summarized ROC (SROC) curve was also proposed for meta-analysis, which has the same properties as the ROC. However, the cost/benefit points of SROC are not obtained by varying the decision threshold as in the ROC. We refer readers to [18]–[20] for further details on SROC.

Although the ROC curve provides a good measure of performance for a single classifier, it is not clear how the individual ROC curves can be used to evaluate the overall classification performance of a multistage classifier. There have not been many studies in this work so far. For example, Hill *et al.* [21] provided an approach based on Boolean rules which uses the Kronecker product of classification results obtained from each individual classifiers to compute the joint and conditional performance matrices and hence the overall ROC of multiple classification systems. However, the approach is not useful in cases of multistage classifiers trained to classify a particular object without knowing the individual ROC of each classifiers. This motivates us to look for a fully or partially new way to evaluate the performance of the multistage classifiers.

In this work, we present our analysis of the ROC curve for a type of multistage classification systems aimed at detecting "positive" data points, as shown in Fig 1. Specifically, in each stage, the individual classifier focuses only on positive points (+1) which are then fed to the next stage. As the result, the number of negative points (-1) will be decreased gradually through the stages of the system. The contribution of the paper is three-fold: 1) providing the range of probable prediction, (2) introducing two new ROC models, and 3) experimenting the proposed ROC models on a recently introduced multistage system for epileptic spike classification from electroencephalogram data.

The paper is organized as follows. Our proposed ROC

978-1-5386-2896-6/17/$31.00 ©2017 IEEE

models are presented in Section II. Experiments and results are shown in Section III, and Section IV concludes the paper.

II. Theory

A. ROC on multistage classification systems

In single classifiers, the ROC curve for binary classification is a plot comprised of cost/benefit points $(1-\text{SPE}_\theta, \text{SEN}_\theta)$ providing the range of probable prediction, where

$$\text{Sensitivity: } \text{SEN}_\theta = \frac{\text{TP}}{\text{TP + FN}}, \tag{1}$$

$$\text{Specificity: } \text{SPE}_\theta = \frac{\text{TN}}{\text{FP + TN}}. \tag{2}$$

Above, TP and FP denote the number of correctly and incorrectly identified samples, TN and FN denote the correctly and incorrectly rejected samples, respectively. The ROC curve increases from 0 to 1, because the two evaluation metrics, SEN and $(1-\text{SPE})$, have to increase or decrease together when the decision threshold θ is varied (the higher the threshold, the lower the sensitivity and the higher the specificity, as shown in Fig 2).

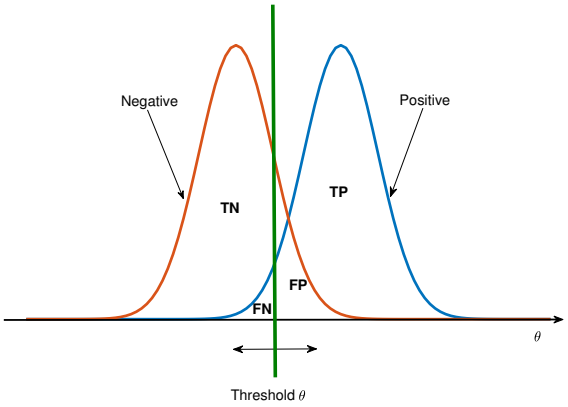

Fig. 2: The decision threshold θ allows us to separate the measurable space into two sub-spaces representing two classes of the dataset.

The notion of the ROC curve of such a single classifier as above can also be generalized to the overall ROC curve of a multistage classification system, for the same purpose of evaluating the performance of the system. In other words, the true positive rate (TPR) and the false positive rate (FPR), which are respectively equivalent to SEN and $1-\text{SPE}$, are needed to draw the overall ROC curve. In this case, the TPR and FPR are determined by several thresholds with different prediction distributions. This then gives rise to another problem: the TPR and FPR may not change together in such a way that leads to a non-increasing ROC curve. Therefore, the overall ROC curve may not be suitable for cost-and-benefit analysis. We propose a new method to estimate the overall ROC curve, which is composed of two steps: 1) determining the range of probable prediction and 2) fitting the ROC model.

B. Determining the range of probable prediction

Theorem 1. *The probable cost/benefit points of the multistage classification system come from the different stage decision thresholds.*

Proof.
Assume that we have a dataset $\mathbf{X} = \mathbf{X}_0 \cup \mathbf{X}_1$ with $\mathbf{X}_0 \cap \mathbf{X}_1 = \{\emptyset\}$. $\mathbf{X}_0$ represents the first class including N_0 data points, while the second class is provided by $\mathbf{X}_1$ with N_1 points. Let $N = N_1 + N_0$. We would like to separate the entire dataset into two clusters. Without loss of generality, let's denote $\mathbf{X}_1$ be positive $(+1)$ set that we are interested in and the negative set (-1) is $\mathbf{X}_0$.

A multistage classification system, $\mathbf{A}$, is then formed from the k individual classifiers, $\mathbf{A}_i$ as $\mathbf{A} = \mathbf{A}_1 \circ \mathbf{A}_2 \circ ... \circ \mathbf{A}_k$, where each $\mathbf{A}_i$ is associated with a decision threshold θ_i.

Denote by $(\mathcal{E}_i, \mathbb{F}_i)$ the measurable space which is called image of the data space $\mathcal{X}$ via the transform $\mathbf{A}_i$.

$$\mathbf{A}_i : \mathcal{X} \to \mathcal{E}_i.$$

1) Stage 1: $\mathbf{A}_1$
For all $x_i \in \mathbf{X}$, taking the transform $\mathbf{A}_1$ of the point corresponding to θ_1 yields

$$y_i = \mathbf{A}_1(x_i|\theta_1) = \text{sign}(\mathbf{A}_1(x_i) - \theta_1). \tag{3}$$

Let $\mathbf{Y} = \{y_1, y_2, ..., y_N\}$ be the set of outcomes of $\mathbf{X}$ via $\mathbf{A}_1$. Let us define

$$\begin{aligned}
\mathbf{X}_{\text{TP},\theta_1} &= \{x_i | x_i \in \mathbf{X}_1, y_i = 1, 1 \le i \le N\} \\
\mathbf{X}_{\text{FP},\theta_1} &= \{x_i | x_i \in \mathbf{X}_0, y_i = 1, 1 \le i \le N\} \\
\mathbf{X}_{\text{FN},\theta_1} &= \{x_i | x_i \in \mathbf{X}_1, y_i = -1, 1 \le i \le N\} \\
\mathbf{X}_{\text{TN},\theta_1} &= \{x_i | x_i \in \mathbf{X}_0, y_i = -1, 1 \le i \le N\}
\end{aligned}$$

If we assume that $\mathbf{Y}_{1,\theta_1} = \{x_i | y_i = 1, 1 \le i \le N\}$ represents "positive" results follows the probability density distribution $f_1(t)$ in the measurable space $\mathcal{E}_1$ and $\mathbf{Y}_{0,\theta_1} = \{x_i | y_i = -1, 1 \le i \le N\}$ is considered as the "negative" with the probability density distribution $f_0(t)$, (as shown in Fig 2), then the following properties can be obtained:

$$\begin{aligned}
&\mathbf{X} = \mathbf{Y}_{1,\theta_1} \cup \mathbf{Y}_{0,\theta_1} \quad \text{and} \quad \mathbf{Y}_{1,\theta_1} \cap \mathbf{Y}_{0,\theta_1} = \{\emptyset\}, \\
&\mathbf{Y}_{1,\theta_1} = \mathbf{X}_{\text{TP},\theta_1} \cup \mathbf{X}_{\text{FP},\theta_1}, \\
&\mathbf{Y}_{0,\theta_1} = \mathbf{X}_{\text{FN},\theta_1} \cup \mathbf{X}_{\text{TN},\theta_1}.
\end{aligned}$$

The number of true positive and false positive points are computed as

$$\text{TP}_1 = \text{N}_1 \int_{\theta_1}^{\sup(\theta_1)} f_1(t)dt \tag{4}$$

$$\text{FP}_1 = \text{N}_0 \int_{\theta_1}^{\sup(\theta_1)} f_0(t)dt \tag{5}$$

The set $\mathbf{Y}_{1,\theta_1} = \mathbf{X}_{\text{TP},\theta_1} \cup \mathbf{X}_{\text{FP},\theta_1}$, represented by the right subspace in Fig. 2), is then fed to the second stage, $\mathbf{A}_2$.

2) Stage 2: $\mathbf{A}_2$.

The input of stage 2 is the set of instances predicted as "positive" $\mathbf{Y}_{1,\theta_1}$. Then, similarly, we also obtain a new

probable "positive" set $\mathbf{Y}_{1,\theta_2} = \mathbf{X}_{\text{TP},\theta_2} \cup \mathbf{X}_{\text{FP},\theta_2}$ in which the new values of TP and FP are updated by

$$\begin{aligned}\text{TP}_2 &= \text{TP}_1 \int_{\theta_2}^{\sup(\theta_2)} f_3(t_2)dt_2 \\ &= \text{N}_1 \int_{\theta_1}^{\sup(\theta_1)} f_1(t_1)dt_1 \int_{\theta_2}^{\sup(\theta_2)} f_3(t_2)dt_2 \end{aligned} \tag{6}$$

$$\begin{aligned}\text{FP}_2 &= \text{FP}_1 \int_{\theta_2}^{\sup(\theta_2)} f_2(t_2)dt_2 \\ &= \text{N}_0 \int_{\theta_1}^{\sup(\theta_1)} f_0(t_1)dt_1 \int_{\theta_2}^{\sup(\theta_2)} f_2(t_2)dt_2 \end{aligned} \tag{7}$$

The set is then fed to the next stage, $\mathbf{A}_3$.

3) Stage K: $\mathbf{A}_k$.

From the obtained set of possible positive points from the stage $\mathbf{A}_{k-1}$: $\mathbf{Y}_{1,\theta_{k-1}}$, we have

$$\begin{aligned}\text{TP}_k &= \text{TP}_{k-1} \int_{\theta_k}^{\sup(\theta_k)} f_{2k-1}(t_k)dt_k \\ &= \text{N}_1 \int_{\theta_1} \int_{\theta_2} \cdots \int_{\theta_k} f_1(t_1)f_3(t_2).\cdots f_{2k-1}(t_k)\ dt_1dt_2...dt_k \end{aligned} \tag{8}$$

$$\begin{aligned}\text{FP}_k &= \text{FP}_{k-1} \int_{\theta_k}^{\sup(\theta_k)} f_{2k-2}(t_k)dt_k \\ &= \text{N}_1 \int_{\theta_1} \int_{\theta_2} ... \int_{\theta_k} f_0(t_1)f_2(t_2)...f_{2k-2}(t_k)\ dt_1dt_2...dt_k \end{aligned} \tag{9}$$

Thus, the TPR and FPR are given by

$$\begin{aligned}&\text{TPR}(\theta_1,\theta_2,...,\theta_k) = \frac{\text{TP}_k}{\text{N}_1} \\ &= \int_{\theta_1} \int_{\theta_2} ... \int_{\theta_k} f_1(t_1)f_3(t_2)...f_{2k-1}(t_k)\ dt_1dt_2...dt_k \end{aligned} \tag{10}$$

$$\begin{aligned}&\text{FPR}(\theta_1,\theta_2,...,\theta_k) = \frac{\text{FP}_k}{\text{N}_0} \\ &= \int_{\theta_1} \int_{\theta_2} ... \int_{\theta_k} f_0(t_1)f_2(t_2)...f_{2k-2}(t_k)\ dt_1dt_2...dt_k \end{aligned} \tag{11}$$

Recalling that $f_i(t)$ is a probability density distribution, hence

$$\int_{\inf(t)}^{\sup(t)} f_i(t)dt = 1, \tag{12}$$

and variables $\{t_1, t_2, ..., t_k\}$ are independent, we then have

$$\begin{aligned}&\lim_{(\theta_1,\theta_2,...,\theta_k)\to(\sup(\theta_1),\sup(\theta_2),...,\sup(\theta_k))} \text{TPR}(\theta_1,\theta_2,...,\theta_k) = 0 \\ &\lim_{(\theta_1,\theta_2,...,\theta_k)\to(\sup(\theta_1),\sup(\theta_2),...,\sup(\theta_k))} \text{FPR}(\theta_1,\theta_2,...,\theta_k) = 0 \\ &\lim_{(\theta_1,\theta_2,...,\theta_k)\to(\inf\theta_1),\inf(\theta_2),...,\inf(\theta_k))} \text{TPR}(\theta_1,\theta_2,...,\theta_k) = 1 \\ &\lim_{(\theta_1,\theta_2,...,\theta_k)\to(\inf(\theta_1),\inf(\theta_2),...,\inf(\theta_k))} \text{FPR}(\theta_1,\theta_2,...,\theta_k) = 1 \end{aligned}$$

□

C. Fitting the ROC curve

As mentioned above, we may not directly produce the overall ROC curve for the multistage classification system from the set of probable cost/benefit points. It is therefore needed to estimate an increasing function presenting the overall ROC curve. This leads to the following optimization problem:

$$\begin{aligned}\underset{f}{\text{minimize}} \quad & \sum_{i=1}^{\text{N}} \text{distance}(f, \text{P}_i) \\ \text{subject to} \quad & f : [0,1] \to [0,1] \\ & \forall u_1, u_2 \in [0,1], \\ & f(u_2) \geq f(u_1) \Leftrightarrow u_2 \geq u_1. \end{aligned} \tag{13}$$

where $\text{P}_i(\text{SEN}_i, 1 - \text{SPE}_i)$ represents a cost/benefit point in the measurable space.

Since the individual ROC curve is usually a nonlinear function except the case of the random effect model, we are interested in mapping the ROC space into a new vector space via a nonlinear transform $h : \mathbb{R}^2 \to \mathbb{R}^2$ in order to fit the ROC model using a common linear regression method. This leads to a new form of problem (13) which can be solved more easiliy in the (D,S) space by linear least-square or robust regression [22]:

$$\text{D} = a\,\text{S} + b \tag{14}$$

$$\underset{a,b}{\text{minimize}} \sum_{i=1}^{\text{N}} \big(\text{D}_i - a\,\text{S}_i - b\big)^2, \tag{15}$$

where (D_i, S_i) is the image of P_i under the transform h. After that, the estimated ROC curve representing the trade-off between TPR and FPR can easily be generated.

We investigate two types of distribution for this task, including the Gaussian distribution,

$$\text{f}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-m)^2}{2\sigma^2}}, \tag{16}$$

and the logistic distribution,

$$\text{f}(x) = \frac{e^{-\frac{x-m}{\sigma}}}{\sigma\big(1 + e^{-\frac{x-m}{\sigma}}\big)^2}. \tag{17}$$

These attempts are motivated by the fact that the error distribution tends to converge to the Gaussian distribution, thanks to the central limit theorem. Therefore, the cost/benefit points $\{\text{P}_i\}$ in the ROC space follow a straight line in the (D,S) space (see Fig. 4). Also, as seen in Fig. 3, the standard logistic distribution is close to the Gaussian distribution.

Taking the inverse CDF of the logistic and Gaussian distributions, namely *logit* and *probit* functions respectively, we have

$$\text{logit}(x) = \ln\big(\frac{x}{1-x}\big), \tag{18}$$

$$\text{probit}(x) = \sqrt{2}\text{erf}^{-1}(2x - 1), \tag{19}$$

where the Gaussian error function is defined by

$$\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt.$$

To draw a common line, as shown in Fig. 4, the relationship

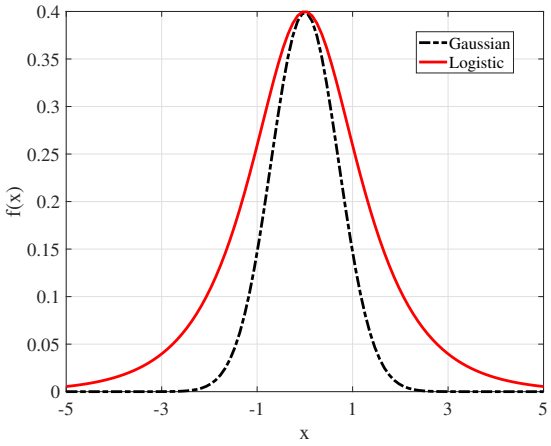

Fig. 3: Comparing PDF of standard logistic distribution and normal distribution.

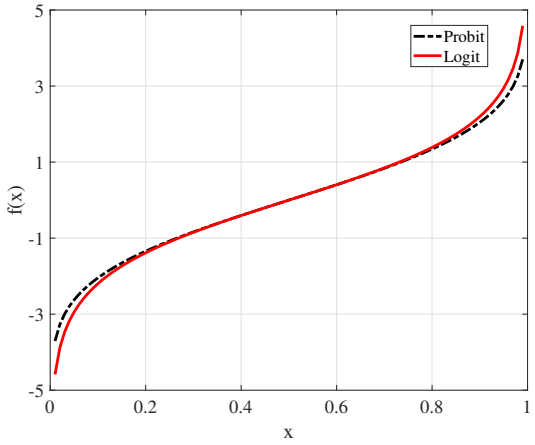

Fig. 4: Comparing inverse CDF of standard logistic distribution and normal distribution.

between the two functions is obtained as

$$\text{logit}(x) \approx \frac{2\sqrt{2}}{\sqrt{\pi}}\text{probit}(x),\ \forall x \in [0,1]. \tag{20}$$

Consequently, the method based on the logistic function should be considered. If the measurement follows the logistic distribution, the results tend to be similar to that of the method based on the Gaussian distribution. This method was employed to combine independent studies of diagnostic clinical testing [18]–[20]. In a nutshell, the two methods can be summarized as follows:

1) Method 1: Gaussian distribution-based:
Define

$$\text{D} = \frac{4}{\sqrt{\pi}}\text{erf}^{-1}(2\,\text{SEN} - 1) - \frac{4}{\sqrt{\pi}}\text{erf}^{-1}(1 - 2\text{SPE}), \tag{21}$$

$$\text{S} = \frac{4}{\sqrt{\pi}}\text{erf}^{-1}(2\,\text{SEN} - 1) + \frac{4}{\sqrt{\pi}}\text{erf}^{-1}(1 - 2\text{SPE}). \tag{22}$$

Combining the above resuls with Eq. (14) yields the estimated ROC curve in the ROC space

$$\text{SEN} = \frac{1}{2} + \frac{1}{2}\text{erf}\left(\frac{1+a}{1-a}\text{erf}^{-1}(1 - 2\text{SPE}) + \frac{b\sqrt{\pi}}{4(1-a)}\right). \tag{23}$$

The area under the estimated ROC curve (AUC) is then given by

$$\text{AUC} = \frac{1}{2} + \frac{1}{2}\int_0^1 \text{erf}\left(\frac{1+a}{1-a}\text{erf}^{-1}(1-x) + \frac{b\sqrt{\pi}}{4(1-a)}\right)dx. \tag{24}$$

2) Method 2: Logistic distribution-based:
Define

$$\text{D} = \ln\left(\frac{\text{SEN}}{1-\text{SEN}}\right) - \ln\left(\frac{1-\text{SPE}}{\text{SPE}}\right), \tag{25}$$

$$\text{S} = \ln\left(\frac{\text{SEN}}{1-\text{SEN}}\right) + \ln\left(\frac{1-\text{SPE}}{\text{SPE}}\right). \tag{26}$$

We obtain another estimate of the ROC curve

$$\text{SEN} = \frac{1}{1 + e^{-b/(1-a)}\left(\frac{1-\text{SPE}}{\text{SPE}}\right)^{(1+a)/(1-a)}}. \tag{27}$$

The AUC is then computed as

$$\text{AUC} = \int_0^1 \frac{1}{1 + e^{-b/(1-a)}\left(\frac{1-x}{x}\right)^{(1+a)/(1-a)}}dx. \tag{28}$$

III. Experiments

A. A Multistage Epileptic Spike Classification System

1) Epileptic spikes:
Epilepsy is a chronic disorder of the nervous system in the brain, characterized by epileptic seizures which are due to abnormal and excessive discharges of nerve cells. The diagnosis of epilepsy is typically based on observation of the seizure onset and the underlying cause. Electroencephalogram (EEG) is one of the most accessible tools supporting this observation. From the EEG recordings, one can detect and classify the specific signal patterns representing the body status. Doctors usually inspect the EEG recordings on a computer screen and look for epileptic spikes, which are abnormal patterns of the brain electrical activity. This conventional manual process is not only very tedious and time-consuming, but also subjective since it depends on the expertise and experience of the doctors. Thus, accurate automatic classification of epileptic spikes is desirable.

The EEG data used in our experiments were recorded at the Signals and Systems Laboratory, using the international standard 10-20 system with 19 channels and the sampling rate of 256Hz. The measurements were carried out on 17 patients aged from 6 to 18 years. The training set is composed of recordings from 12 patients, while data from remaining five patients are used for testing, as shown in Tab. I.

2) The Classification System:
The structure of this multistage spike classification system is shown in Fig. 5. It processes input EEG recordings in four main steps as follows

- *Pre-processing*: The system first detects all peaks in the EEG signal, then identifies and removes small peaks, whose amplitude and duration are less than 20ms and

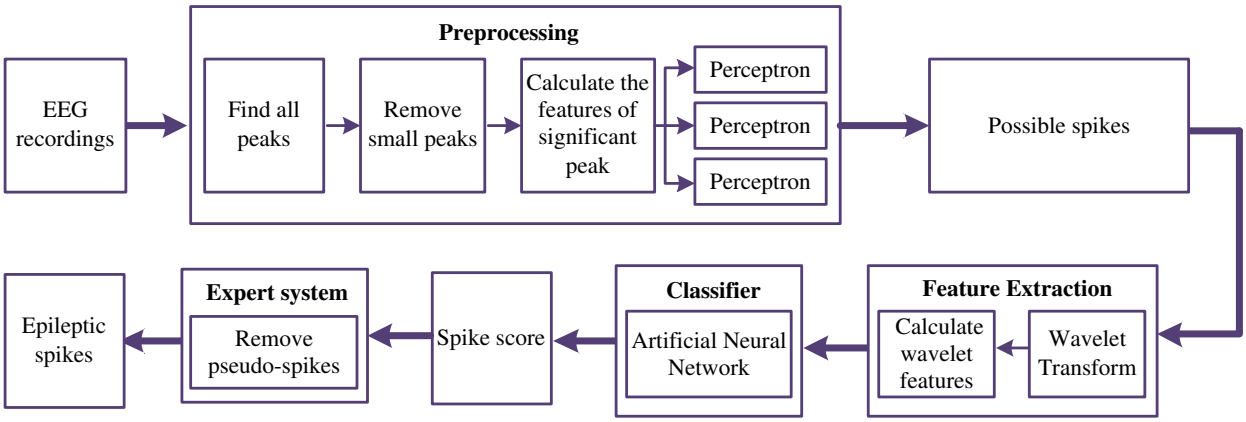

Fig. 5: The multistage system for epileptic spike detection.

TABLE I: EEG Dataset

Training Set			Testing Set		
Case	Duration	Spikes	Case	Duration	Spikes
1	23m57s	8	1	11m24s	16
2	22m25s	635	2	27m13s	1
3	11m24s	6	3	16m16s	351
4	19m21s	8	4	5m31s	12
5	22m0s	4	5	27m37s	19
6	17m49s	22	6	27m37s	9
7	15m26s	2			
8	22m58s	11			
9	20m14s	8			
10	18m53s	5			
11	14m32s	324			
12	17m7s	2			

TABLE II: The classification results

Case	Spikes	TP	FP	FN	TN	SEN(%)	SPE(%)
1	16	14	6262	2	30609	87.5	83.01
2	1	1	839	0	6135	100	87.97
3	351	323	4034	28	32004	92.02	88.8
4	12	12	3391	0	21539	100	86.4
5	19	18	3126	1	10107	94.74	76.37
6	9	9	3674	0	10543	100	74.16

2μV, respectively. After that, for each remaining significant peak, six spike features characterizing durations, amplitudes, and slopes are extracted. These features are then used as inputs for three independent perceptrons which will classify the peaks into two groups: non-spikes and possible spikes.

- *Feature Extraction*: The possible spikes are decomposed by a continuous wavelet transform. The transformed signal corresponding to each scale is then used to calculate seven wavelet features. A feature vector of 35 components for each possible spike, comprised of wavelet features of the wavelet scales from 4^{th} to 8^{th}, is created and used as an input value for classification stage.
- *Classification*: Feature vectors representing possible spikes are fed into a classification model producing a binary class label as its output. The spike score values are in the range $[0, 1]$. A peak is labeled as a spike when its corresponding output score is higher than a certain threshold defined by the model. This stage uses the Artificial Neural Network (ANN) learning model which calculates the probability of a test vector to be assigned to an epileptic spike class.
- *Expert System*: Finally, in order to enhance the accuracy of classification, we applies an heuristic that there are at most two spikes in a duration of 150-350ms to eliminate the pseudo spikes which are located near an epileptic spike.

We refer readers to [23] for more details.

B. Experimental Results

Results obtained from our experiments with test data are presented in Tab. II and plotted in Fig. 6. Tab. II provides quantitative statistics of the best spike classification results. As seen from the results, SEN of the system is always high ($> 87.5\%$), while the SPE varies from patient to patient but is still reasonably good, with value varying from 74.16% to 88.8%. Fig. 6 shows the range of probable results of the classification by varying two decision thresholds of the ANN and the expert system. We also wish to draw an ROC curve characterizing the systems's performance. Unfortunately, the cost/benefit points do not follow an increasing function as in the interval $[0, 1]$ as the classical ROC curve should do, that make using it inappropriate.

Then, our proposed models are used to estimate the ROC curve. The estimated ROCs are shown in the Fig. 7 and their parameters are shown in Tab. III. The two estimated ROCs

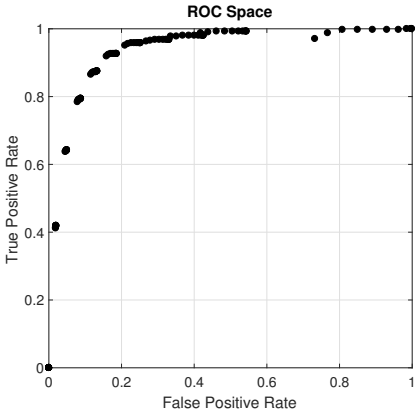

Fig. 6: The cost/benefit points in 2D ROC space.

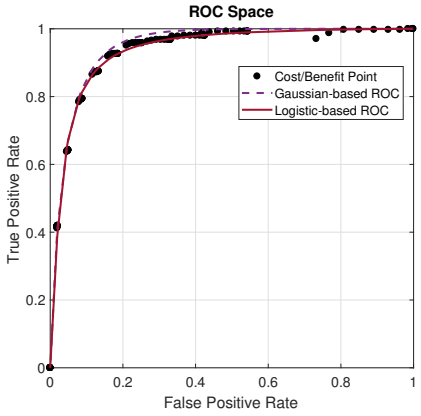

Fig. 7: Estimated ROCs in the multistage epileptic spike classification system.

TABLE III: Results on ROC estimation

Method	a	b	AUC	Error
Gaussian-based	0.24	3.8	0.9490	0.0157
Logistic-based	0.1	3.9	0.9376	0.0093

always follow the cost/benefit points in ROC space. Both ROC estimation methods yield similar results with reasonable errors (< 0.02) and AUC ≈ 0.94. The logistic distribution-based method seems to be flatter while Gausssian distribution-based method gives higher AUC, but the difference is insignificant.

IV. Conclusion

We propose a new approach to evaluating the performance of multistage classification systems with two new ROC estimation models. The models respectively use Gaussian and logistic distributions to fit the ROC curve, motivated by the standard distribution of errors in detection and estimation theory. The proposed methods have been validated by mathematical means as well as by experiments with an automatic epileptic spike classification system using data from real patients. Experimental results, though still preliminary, indicate the great potential of our proposed approach.

References

[1] S. Raghunathan, A. Jaitli, and P. P. Irazoqui, "Multistage seizure detection techniques optimized for low-power hardware platforms," *Epilepsy & behavior*, vol. 22, pp. S61–S68, 2011.

[2] H. S. Liu, T. Zhang, and F. S. Yang, "A multistage, multimethod approach for automatic detection and classification of epileptiform EEG," *IEEE Transactions on Biomedical Engineering*, vol. 49, no. 12, pp. 1557–1566, 2002.

[3] H. Meng and N. Bianchi-Berthouze, "Naturalistic affective expression classification by a multi-stage approach based on hidden markov models," *Affective computing and intelligent interaction*, pp. 378–387, 2011.

[4] Z. Xiao, E. Dellandrea, W. Dou, and L. Chen, "Multi-stage classification of emotional speech motivated by a dimensional emotion model," *Multimedia Tools and Applications*, vol. 46, no. 1, p. 119, 2010.

[5] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[6] R. Herbei and M. H. Wegkamp, "Classification with reject option," *Canadian Journal of Statistics*, vol. 34, no. 4, pp. 709–721, 2006.

[7] J. Gama and P. Brazdil, "Cascade generalization," *Machine Learning*, vol. 41, no. 3, pp. 315–343, 2000.

[8] C. N. Silla Jr and A. A. Freitas, "A survey of hierarchical classification across different application domains," *Data Mining and Knowledge Discovery*, vol. 22, no. 1-2, pp. 31–72, 2011.

[9] J. Huang, *Performance measures of machine learning*. University of Western Ontario, 2006.

[10] T. Fawcett, "An introduction to ROC analysis," *Pattern recognition letters*, vol. 27, no. 8, pp. 861–874, 2006.

[11] K. Hajian-Tilaki, "Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation," *Caspian journal of internal medicine*, vol. 4, no. 2, p. 627, 2013.

[12] P. Martínez-Camblor, C. Carleos, and N. Corral, "General nonparametric ROC curve comparison," *Journal of the Korean Statistical Society*, vol. 42, no. 1, pp. 71–81, 2013.

[13] V. I. de Carvalho, A. Jara, T. E. Hanson, M. de Carvalho *et al.*, "Bayesian nonparametric ROC regression modeling," *Bayesian Analysis*, vol. 8, no. 3, pp. 623–646, 2013.

[14] A. Erkanli, M. Sung, E. Jane Costello, and A. Angold, "Bayesian semi-parametric ROC analysis," *Statistics in medicine*, vol. 25, no. 22, pp. 3905–3928, 2006.

[15] T. Cai and C. S. Moskowitz, "Semi-parametric estimation of the binormal ROC curve for a continuous diagnostic test," *Biostatistics*, vol. 5, no. 4, pp. 573–586, 2004.

[16] G. Hughes and B. Bhattacharya, "Symmetry properties of bi-normal and bi-gamma receiver operating characteristic curves are described by Kullback-Leibler divergences," *Entropy*, vol. 15, no. 4, pp. 1342–1356, 2013.

[17] E. Hussain, "The bi-gamma ROC curve in a straightforward manner," *Journal of Basic & Applied Sciences*, vol. 8, no. 2, pp. 309–314, 2012.

[18] S. Steinhauser, M. Schumacher, and G. Rücker, "Modelling multiple thresholds in meta-analysis of diagnostic test accuracy studies," *BMC medical research methodology*, vol. 16, no. 1, p. 97, 2016.

[19] M. Korsten, "Application of summary receiver operating characteristics (sROC) analysis to diagnostic clinical testing," *7 Reflections on the future of gastroenterology–unmet needs*, vol. 52, p. 76, 2007.

[20] J. Zamora, V. Abraira, A. Muriel, K. Khan, and A. Coomarasamy, "Meta-DiSc: a software for meta-analysis of test accuracy data," *BMC medical research methodology*, vol. 6, no. 1, p. 31, 2006.

[21] M. E. O. Hill, Justin M. and K. W. Bauer, "Receiver operating characteristic curves and fusion of multiple classifiers." in *Proceedings of the 6th International Conference on Information Fusion*, vol. 2, 2003.

[22] D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant, *Applied logistic regression*. John Wiley & Sons, 2013, vol. 398.

[23] N. T. Anh Dao, N. Linh Trung, N. Van Ly, T. Duc Tan, N. T. Hoang Anh, and B. Boashash, "A multistage system for automatic detection of epileptic spikes," *REV Journal on Electronics and Communications*, 2017 (submitted).