Confident Surgical Decision Making in Temporal Lobe Epilepsy by Heterogeneous Classifier Ensembles

Shobeir Fakhraei[1,2] shobeir@wayne.edu
Hamid Soltanian-Zadeh[2,3] hamids@rad.hfh.edu
Kourosh Jafari-Khouzani[2] kjafari@rad.hfh.edu
Kost Elisevich[4] nskoe@neuro.hfh.edu
Farshad Fotouhi[1] fotouhi@wayne.edu

1. Dept. of Computer Science, Wayne State University, Detroit, MI, USA
2. Image Analysis Lab., Dept. of Radiology, Henry Ford Health System, Detroit, MI, USA
3. CIPCE, School of Elec. and Comp. Eng., University of Tehran, Tehran, Iran
4. Dept. of Neurosurgery, Henry Ford Health System, Detroit, MI, USA

Abstract— **In medical domains with low tolerance for invalid predictions, classification confidence is highly important and traditional performance measures such as overall accuracy cannot provide adequate insight into classifications reliability. In this paper, a confident-prediction rate (CPR) which measures the upper limit of confident predictions has been proposed based on receiver operating characteristic (ROC) curves. It has been shown that heterogeneous ensemble of classifiers improves this measure. This ensemble approach has been applied to lateralization of focal epileptogenicity in temporal lobe epilepsy (TLE) and prediction of surgical outcomes. A goal of this study is to reduce extraoperative electrocorticography (eECoG) requirement which is the practice of using electrodes placed directly on the exposed surface of the brain. We have shown that such goal is achievable with application of data mining techniques. Furthermore, all TLE surgical operations do not result in complete relief from seizures and it is not always possible for human experts to identify such unsuccessful cases prior to surgery. This study demonstrates the capability of data mining techniques in prediction of undesirable outcome for a portion of such cases.**

Keywords-Classification; Confidence-based Classification; Confident Prediction; AUC; Performance Evaluation; Ensemble Methods; Epilepsy; Lateralization; Outcome; Temporal Lobe

I. INTRODUCTION

Data mining techniques have been successfully applied in various biomedical and bioinformatics problems to study complex diseases [1]. In studying epilepsy, availability of several diagnostic methods from multiple sources results in creation of high-dimensional spaces where without the aid of appropriate tools, data analysis and decision making become intricate tasks. In such domains, utilization of data mining tools and techniques could result in substantial benefits.

Epilepsy is a disorder of the brain characterized by an enduring predisposition to generate epileptic seizures and by the neurobiological, cognitive, psychological and social consequences of this condition [2]. The most operated form of localization-related epilepsy is mesial temporal lobe epilepsy (mTLE). Neurosurgical resection of the abnormal brain tissue in patients suffering from mTLE is a practiced method of seizure elimination and reduction. Prior to such operation, focal points of the seizure should be lateralized via a set of examinations.

To lateralize the seizure focus in mTLE patients, several noninvasive clinical attributes are investigated. Once decisive evidence could not be found in patients' noninvasive clinical profiles, extraoperative electrocorticography (eECoG) is required, which is the practice of using electrodes placed directly on the exposed surface of the brain to record electrical activities from the cerebral cortex. Such procedure adds financial burden and further distress to the whole surgical resection process.

A possible way of reducing such requirements is via case based reasoning and learning from previously operated patients with similar assessment parameters. In this paper, application of data mining methods in building a recommender system to reduce the need for eECoG is reported.

The main data mining task in such recommender system is a binary classification of laterality. Lateralization of seizure focus is identifying the side of abnormality in either "right" or "left" side of the brain. However, as explained in the next section, utilizing classical classifiers in such binary classification does not help in reduction of eECoG requirement and a method with higher confident predictions is needed.

Furthermore, due to various possibly unknown reasons, not all surgical resections result in complete relief from seizures. Since the surgery is irreversible, such waste is significant. Currently, there is no method to identify the patients with higher chance of having an undesirable surgical outcome prior to surgery. A recommender system that could flag high risk patients would alert the physicians and surgeons to advance the operation with more caution. In this paper, we report a method of patient scoring to build such a system.

In the following sections, different attributes, preprocessing stages, and classification tasks are explained. Confident prediction measures and ensemble methods that improve classification performance are described.

II. STUDY METHOD

For the lateralization task, standard data mining preprocessing techniques are applied. A subset of most discriminative features is selected and missing values are treated with imputation. Several classifiers are applied and their performances are evaluated based on a proposed measure of prediction confidence. Furthermore, seven different ensemble functions are applied to prediction combinations and their effectiveness in improving the confidence measure are investigated.

For surgical outcome prediction, combination of three class membership scores is proposed and its performance reported. In the following subsections, the proposed confident prediction performance measure and the ensemble method are described.

A. Performance Evaluation

In our problem, the goal of the recommender system is to reach sufficient decision confident to ultimately eliminate the need for further investigations such as eECoG. In other words, such system has to be reliable enough to replace the well established eECoG investigation. Consequently, the domain has very low tolerance for invalid predictions from the system. In contrast, error is an inevitable part of classification, and while training is done to minimize the overall error, having an error free prediction in real world applications is virtually impossible.

In such critical domains that require high level of reliability, a method that increases the decision confidence is preferred over systems with high overall accuracy. A confidence-based classification system would only provide predictions for cases with achievable decision confidence above a certain threshold. Other cases would be considered not decidable and hence rejected by the classifier for a decision [3].

As an extreme example, in our scenario, a classifier that could always make a valid prediction on 10% of cases and has no predictions for the other 90% is preferred over a classifier with 95% accuracy that provides predictions for all cases. Since it is impossible to find cases where the 5% error has occurred in the later classifier, all of the predictions hold a chance of being invalid and the eECoG requirement could not be eliminated. However, the former classifier virtually never makes a mistake for the predicted 10%, and it could potentially replace the need for eECoG operation in 10% of the cases with no further investigations.

To build such a classifier, the task of binary classification is changed to two overlapping recognition tasks. Instead of having a boundary that partitions the feature space into "left" or "right" or positive and negative (Figure 1A), two boundaries are established to partition the space into "left", "right", and "undecided" regions (Figure 1B).

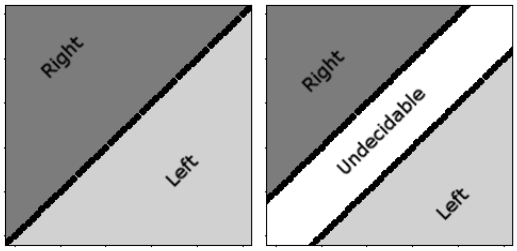


Figure 1. Partitioning the feautre space into two possible prediction regions in binary classification (A), compared to partitiontioning the space into three regions where there is an "undecided" region in the middle (B).

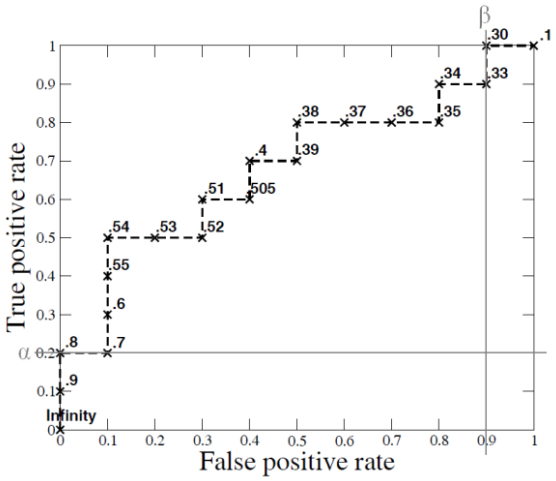
In this method, decision boundaries should be chosen such that accuracy of the predictions in "left" and "right" which we call "confident prediction" regions are above a user mandated threshold. Classifiers preference is based on having smaller "undecided" regions. In this paper which covers our specific domain we have chosen the extreme threshold of 100% accuracy for decisions. Although such accuracy could not be guaranteed for unseen cases, it is the highest achievable confidence based on training data.

While aspects of choosing the correct decision boundaries is not the purpose of this paper, we have proposed a method of comparing the classifiers based on their confident predictions limits using receiver operating characteristic (ROC) curves. Some of the related efforts that consider the confidence in classifier design are reported in [4, 5].

ROC curves are the preferred ways of comparing classifier performance over accuracy for several reasons such as presence of class imbalance in datasets [6]. Using a ranking or scoring list of predictions such as probability of class membership, ROC curves could be plotted by changing decision boundaries from $-\infty$ to $+\infty$ [7]. Area under the ROC curve (AUC) is a common way of numerical quantification of the curves. However, for the purpose of measuring the confident predictions, calculating the AUC is not the best method and other measures are required.

Figure 2 illustrates an ROC curve generated based on an example provided in [7]. In this example, the third instance is the false positive with the highest incorrect positive prediction score. In this case, it is apparent that a confident decision boundary or threshold should be greater than 0.7, since any threshold less than that will include the incorrect decision of the third instance. The threshold could be chosen more conservatively, but the score of the third instance is the lower limit for the confident decision thresholds. In the ROC plot of Figure 2, the third instance corresponds to the point where the curve is deviated from the vertical axis. We refer to this point as α which is shown by a horizontal gray line and is the lower limit for all confident decision thresholds for this classifier. Respectively, on the other side, the nineteenth

instance is the worst false negative and is the upper limit for all confident decision thresholds for this classifier. The limit is shown by a vertical gray line on Figure 2 and is referred to as β.



Inst#	Class	Score
1	p	0.9
2	p	0.8
3	n	0.7
4	p	0.6
5	p	0.55
6	p	0.54
7	n	0.53
8	n	0.52
9	p	0.51
10	n	0.505

Inst#	Class	Score
11	p	0.4
12	n	0.39
13	p	0.38
14	n	0.37
15	n	0.36
16	n	0.35
17	p	0.34
18	n	0.33
19	p	0.3
20	n	0.1

Figure 2. The ROC curve created by thresholding a test set. The table shows twenty data points and the scores assigned to each by a scoring classifier. The graph shows the corresponding ROC curve with each point labeled by the threshold that produces it [7].

The instances whose score is greater than α or less than β will be the maximum number of instances that the system could provide a confident prediction for and the instances between α and β would fall in the undecided region. We refer to such quantity extracted from the ROC curve as the confident-prediction rate (CPR) [8] and write it as:

$$CPR = \frac{(\#samples \mid score>\alpha)+(\#samples \mid score<\beta)}{Total\ number\ of\ samples} * 100 \quad (1)$$

where α is the score of the instance of negative class with the highest score and β is the score of the instance from positive class with the lowest score, when the instances are scored based on prediction of their membership to the positive class. On the ROC curve α is the point where the curve deviates from the vertical axis and β is the point where the curve deviates from the top horizontal axes.

Real classification thresholds should be established by having two sets of unseen instances as validation and test sets, but the optimistic estimate is the CPR calculated from the ROC curve. In other words, CPR is the upper limit for the confident prediction of a classifier, but the real number of instances with confident prediction depends on how conservatively the classification thresholds are chosen and how extreme they are from α and β.

It is also interesting to note that CPR is not correlated with AUC. As an example, the ROC of the logistic regression (LR) and random forests (RF) classifiers used for the lateralization of the seizure focus in our scenario are shown in Figure 3. It could be seen that although LR generates the AUC of 0.985 and RF results in AUC of 0.968, CPR of RF is higher than that of LR. This is due to the ROC in LR deviating from the vertical axis very early and setting α too high, which results in CPR of 44.3% for LR comparing to 64.6% for RF.

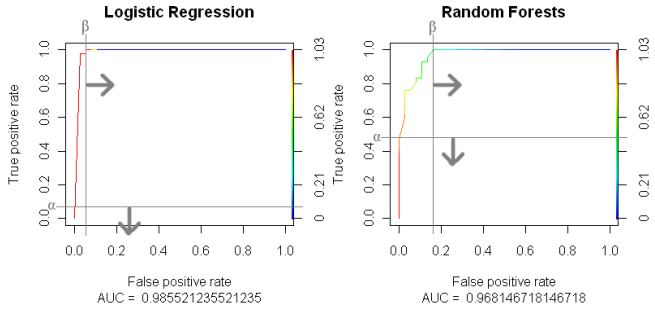


Figure 3. Receiver operating characteristic (ROC) curves of logistic regression vs. random forests classifiers in the study. α and β thresholds are shown by gray lines on the charts.

Our goal in this study is to find classification methods that maximize the CPR. Since such methods could potentially result in elimination of the eECoG requirements prior to resection for a portion of the cases.

B. Heterogeneous Classifier Ensemble

It is known that ensemble of classifiers that has independent errors improve the overall accuracy of the classifiers. Lowering the chance of getting stuck in local optima, reducing the risk of choosing the wrong classifier, and expanding the space of representable functions are the main reasons for such phenomena [9].

Using the proposed measure of prediction confidence, we show that a heterogeneous ensemble of classifiers improves prediction confidence. Heterogeneous ensemble of classifiers is when the classifiers participating in the ensemble are not of the same type. This method might also be referred to as consensus learning.

Six classifiers are applied to the lateralization task with preoperative data of patients to assess the possibility of predicting the side of abnormality. These data exclude the

invasive eECoG measurements. Several ensemble functions are evaluated and their performance improvements over single classifiers are reported in terms of AUC and CPR.

III. DATASET

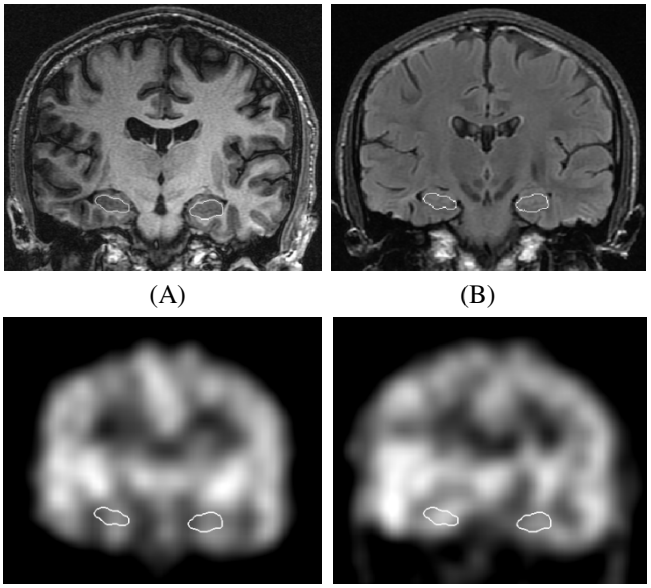
Various clinical attributes of mTLE patients from various sources and subsystems are gathered over the past several years in an integrated database at the radiology research department of Henry Ford Health System in Detroit Michigan.

The attributes includes descriptive electrographic data (EEG), images, Wada test, semiology, risk factors underlying the condition, neuropsychological profiles, locations of surgery, pathology and outcome according to the Engel classification (Class-I: Free of disabling seizures, Class-II: Rare disabling seizures, Class-III: Worthwhile improvement, Class-IV: No worthwhile improvement).

Evaluating with multiple feature selection methods, imaging, EEG, Wada and Neuropsychological attributes were the most discriminative features for lateralization of the seizure focus [10]. Description of these attributes and the preprocessing steps to extract quantitative values from them are provided in this section.

A. Imaging Features

Imaging features were generated using the hippocampi outlines. A domain expert outlined all hippocampal contours on coronal slices of T1-weighted images using in-house software and a previously established protocol (Figure 4A) [11]. These were then verified by another expert.



(A) (B) (C) (D)

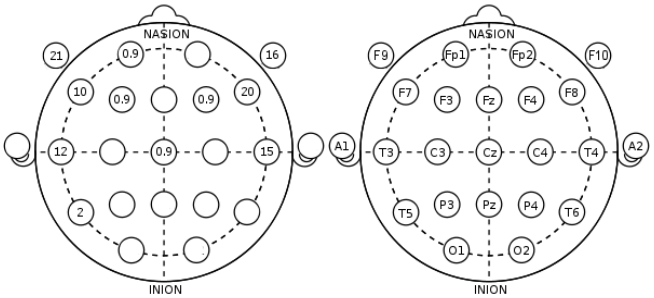
Figure 4. Manual segmentation of the hippocampi in a representative coronal T1-weighted MR image (A) and its map on the FLAIR MRI (B), interictal SPECT (C), and ictal SPECT (D).

Imaging features were extracted as follows using previously established methods [12, 13]. The T1-weighted and fluid-attenuated inversion recovery (FLAIR) magnetic resonance (MR) image sets were co-registered using a rigid registration technique based on mutual information [14]. Similarly, ictal and interictal single photon emission computed tomography (SPECT) image sets were co-registered to the T1-weighted image set. To increase the accuracy of co-registration, non-brain tissues in T1-weighted and FLAIR MR images were eliminated using brain extraction tool (BET) [15]. The manually segmented regions of interest (ROIs) were mapped onto the FLAIR MR and SPECT image sets using the registration parameters (Figure 4B-D).

Four sets of features were extracted from each hippocampal ROI: mean and standard deviation of the FLAIR MR signal intensity, wavelet transform-derived energy, volumetry, and SPECT ictal-interictal mean difference. The final value for each feature was expressed as a ratio of measured values of the two hippocampi for the first three features and difference of such for the last feature.

B. Electroencephalography (EEG) Features

Dataset includes descriptive noninvasive electrographic features. Ictal onset locations and three most predominant interictal localities of sharp and slow waveforms are provided. Ictal onset locations were integrated into one feature indicating the probability of focal epileptogenicity in the right temporal lobe. Figure 5B demonstrates the surface electrode configuration of the EEG recordings. The frequency percentage of the two most dominant interictal sharp wave activities in each location for all patients in the dataset are shown in Figure 5A.



(A) (B)

Figure 5. The surface electrode configuration of the EEG recordings (B), and the frequency percentage for sharp wave interictal activities in the entire dataset (A)

C. Neuropsychological Features and Wada Test

Neuropsychological profile of the patients include pre- and post-operative measurement through Boston naming test (BTN), Wechsler memory scale (WMS), Rey-Osterrieth complex figure test (ROCF), California verbal learning test, and intelligence quotient (IQ) test. Quantitative measures of

patients' verbal and non-verbal memory are recorded and stored in the database.

Wada test which is also known as the "intracarotid sodium amobarbital procedure" (ISAP), is used to establish cerebral language and memory representation of each hemisphere. Laterality of language dominance in cerebral hemispheres, and the number of correctly recalled items after left and right carotid injections, is also stored in the database.

IV. Patients Cohorts

In an effort to maintain ground truth for the lateralization, only those cases with surgical outcomes of "free of disabling seizures (Engel class-I)" were considered, as they confirm a definite lateralization by human expert.

In the study of lateralization, 79 patients with Engel class-I outcome were selected (31 males, 48 females) with 197 medical features. The patients have an average age of 38y (SD = 12.2). Seizure focus was found to be on the left side in 43 patients and the right side in 36 patients. In 46 patients, resection proceeded standard noninvasive evaluations, whereas 33 patients required eECoG.

In this cohort, appropriate values were not recorded for EEG features in 21% of cases, for Wada studies in 31%, for SPECT imaging features in 35%, and for FLAIR and volumetric imaging in less than 10% of cases in the dataset. The missing values of the remaining features were found in about 20% of cases on average.

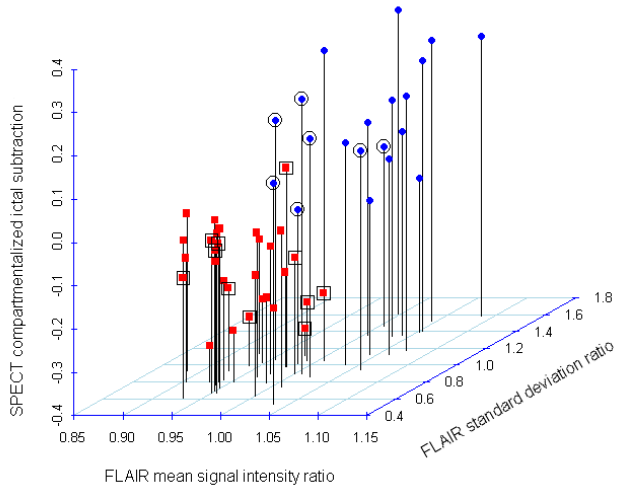


Figure 6. Patients in scatter plot of FLAIR standard deviation ratio, FLAIR mean signal intensity ratio and SPECT compartmentilized ictal subtraction. Patients with abnormality in their right side are shown with circles and the ones with left side abnormality with squares. Phase II patients are outlined. Cases with a missing value in either of the attributes are removed.

We refer to patients whom standard noninvasive evaluations were sufficient for their lateralization as phase I patients and those who required eECoG as phase II patients. Since the side of abnormality in phase I patients could be found by human experts using only the noninvasive studies, a recommender system could be built solely based on phase II patients to address our goal of reducing eECoG requirements. However, such a decision support system should also be able to predict the side of abnormality in phase I patients to increase the confidence of experts decision. Furthermore, limited number of phase II patients in the dataset reduces the classifier's learning power and increases the risk of over fitting. Therefore, both phase I and phase II patients have been included in the learning process of the classifiers.

FLAIR mean and standard deviation ratios and SPECT compartmentalized subtraction of phase I and phase II patients are shown in Figure 6. It is seen that phase I patients populate the spaces that helps the classification task. Cases with missing values in either of the attributes are removed for visualization purpose. Such perfect separation of patients according to the laterality of their seizure focus is not possible when all patients (with missing values) are considered.

V. Experimental Results

The dataset contains a relatively high number of attributes and a limited number of patients. Furthermore, due to the presence of missing values in almost all attributes, inclusion of more features in the classification task requires additional missing value treatments. Consequently, attributes were ranked and a feature subset were selected using the methods described in [16, 17]. The selected feature subset was verified by Relief [18] via separate investigations resulting in a comparable subset.

Most of patients have at least one feature with a missing value and thus elimination of patients with missing values was not reasonable. Therefore, in this stage, missing values were imputed with neutral values. (1 for attributes representing a ratio and 0 for attributes that contained subtractions of two values).

Six well known classifiers were utilized to lateralize the side of abnormality in patients. The codes to support different stages of the experiments were implemented in Java, WEKA, and R. Using the leave-one-out method, a relative score of "right" class membership was generated for each sample. Samples were then sorted according to this probability and the corresponding ROC curve was generated. CPR and AUC measures were computed based on the ROC curve. The classifiers included in this study were naïve Bayes (NB), support vector machine (SVM), 3-nearest neighbors (3NN), multilayer perceptron (MLP), logistic regression (LR), and random forests (RF).

AUC and CPR measures of each classifier are reported in Table I. Naïve Bayes generated the best results according to both measurements. However, as previously mentioned, for the other classifiers AUC is not correlated with CPR. For example in logistic regression classifier where the AUC is 0.986 and CPR is 44.3% while multi-layer perceptron results in lower AUC of 0.978 with higher CPR of 72.2%. In our scenario, despite the higher AUC of LR, MLP is the preferred classifier.

TABLE I. PERFORMANCE OF THE INDIVIDUAL CLASSIFIERS FOR LATERALIZATION OF FOCAL EPILEPTOGENICITY.

Classifier	AUC	CPR
Naïve Bayes	0.993	84.8%
Support Vector Machines	0.959	36.7%
Multi Layer Perceptron	0.978	72.2%
3-Nearest Neighbors	0.964	43.0%
Logistic Regression	0.986	44.3%
Random Forests	0.968	64.6%

For the next experiment, class membership probabilities of the samples are combined using eight different ensemble functions to investigate whether such heterogeneous ensemble increases the performance over single classifiers. Each sample was assigned ensemble probabilities corresponding to arithmetic and geometric means, median, maximum and, minimum of the probabilities assigned to them by all six classifiers. Two other ensemble functions referred to as "optimistic ensemble (OE)" and "pessimistic ensemble (PE)" are proposed and used in this study. As shown below, OE is the most extreme probability toward 0 or 1 and PE is the most conservative probability which is closest to 0.5. While OE takes a more risky approach, using PE generates a conservative prediction.

$$OE = \{P_R(c,s) \mid |P_R(c,s) - 0.5| = max(|P_R(\forall c,s) - 0.5|)\}$$
$$\mathrm{PE} = \{\mathrm{P_R(c,s)} \mid |\mathrm{P_R(c,s)} - 0.5| = \min(|\mathrm{P_R}(\forall \mathrm{c,s}) - 0.5|)\} \quad (2)$$

where c is a classifier, s is a sample and $\mathrm{P_R(c,s)}$ is the score (or probability) of right sided abnormality given by c to s.

AUC and CPR measurements generated using the ensemble functions are summarized in Table II. Although there is no significant improvement in the AUC of classifications, it could be seen that on average, CPR measures are improved. While the median ensemble function generates the best performance that improves the most superior single classifier NB from 84.8% to 88.6%. Using the ensemble approach 81.8% of the phase II patients that required eECoG investigations could be potentially lateralized using data mining techniques based on non-invasive methods, where only 6.5% of patients that were lateralized by human experts using non-invasive measures lay on the undecided region. Of course, traditionally practiced procedures such as eECoG could be followed for patients rejected by the system for a decision.

For the surgical outcome prediction, the patients are categorized into two classes of free of disabling seizures (Engel class-I) and patients with some post-surgical seizures (Engel class-I to class-IV). 108 patients were included in this study, having similar amount of missing values to the patient populations in the previous study. The recorded clinical attributes demonstrated no significant discriminative power between the post-operative seizure-free and seizure-bearing patients.

TABLE II. PERFORMANCE OF HETEROGENOUS ENSEMLBE FUNCTIONS FOR LATERALIZATION OF FOCAL EPILEPTOGENICITY.

Ensemble Function	AUC	CPR
Mean	0.985	78.5%
Median	0.993	88.6%
Maximum	0.959	44.3%
Minimum	0.981	72.2%
Geometric Mean	0.985	78.5%
Pessimistic Ensemble	0.974	69.6%
Optimistic Ensemble	0.973	45.6%

However, scoring the patients as being post-operative seizure-bearing according to asymmetry in the hippocampus volume of the patients results in a higher CPR of 13.9% relative to other features. More interestingly, when instances are scored based on the variance of the lateralization predictions by six different classifiers in the previous study, the CPR of 8.4% is achieved for outcome prediction. This measure demonstrates the lack of consensus among the classifiers for laterality. Another measure extracted from lateralization with multiple classifiers is the average distance of the lateralization predictions from 0.5 which shows how confidently the instance was lateralized. When instances are scores based on this measure, a CPR of 7.5% is achieved for outcome prediction.

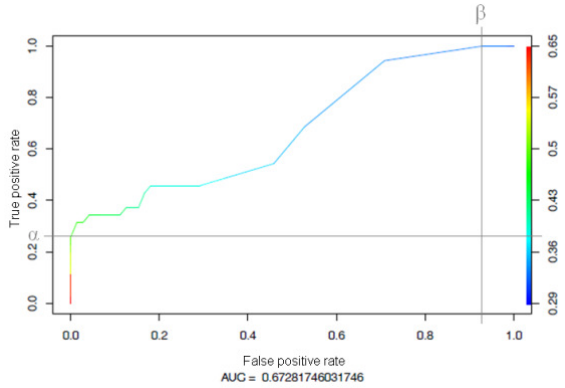


Figure 7. Receiver operating characteristic (ROC) curve of surgical outcome prediction based on scores generated using variations in the lateralizaion predictions, distance of the predictions from 0.5 and, volume assymetry of the hippocampus in patients.

After normalization of these three measures, each instance was scored based on average scores of the three as an indication of how likely the patients are to be seizure-bearing after surgery. The ROC curve generated using this ensemble scoring is plotted in Figure 7. Although the AUC is 0.67 which is not considerably high, CPR is 23.2%. More interestingly, 32.4% of the post-operative seizure-bearing patients lay inside the confident prediction region suggesting that near one-third of the patients who did not improve significantly after the surgery could be identified by this system.

VI. Discussion and Conclusion

In the paper, a limitation of the traditional classification performance measures in medical domains with low tolerance for invalid predictions is highlighted. Confident prediction rate (CPR) which is a method to measure the limit of confident predictions is proposed based on ROC curves. With comparing multiple ensemble functions, it is been shown that a heterogeneous ensemble of classifiers improves the CPR measure.

This ensemble method is applied to lateralization and surgical outcome prediction in temporal lobe epilepsy resection. It is shown that up to 88.4% of the patients could be lateralized based on this system while without the use of such data mining technique, only 58.2% of the patients were lateralized by the domain experts using noninvasive methods.

Surgical outcome of the patients may be predicted on 23.2% of the patients as completely seizure free and seizure bearing. A total of 32.4% of the post-surgical seizure bearing patients may be potentially identified by this system. However, the methods used for outcome prediction requires further investigation since the results should be evaluated by cross validation. Furthermore, although more confident predictions in outcome prediction will assist in easier medical decision making, such low tolerance in error is not of high priority in this case.

Finally, this study shows that classifiers that could optimize CPR are of interest in critical domains such as medicine which has very low tolerance for invalid predictions, and methods that could train classifiers for such a purpose are required.

Acknowledgment

Special thanks to Dr. Romina Shirka and Dr. Toya Malone for assisting with the EEG data collection and Dr. Kenneth Podell for providing access to neuropsychological profiles of the patients. This work was supported in part by NIH grant R01-EB002450.

References

[1] R. Bellazzi and B. Zupan, "Predictive data mining in clinical medicine: Current issues and guidelines," International Journal of Medical Informatics, vol. 77, pp. 81-97, 2008.

[2] C. P. Panayiotopoulos, A clinical guide to epileptic syndromes and their treatment: Springer Verlag, 2010.

[3] C. K. Chow, "On optimum recognition error and reject tradeoff," IEEE Transactions on Information Theory, vol. IT-16, pp. 41-6, 1970.

[4] J. B. Tilbury, W. J. Van Eetvelt, J. M. Garibaldi, J. S. H. Curnsw, and E. C. Ifeachor, "Receiver operating characteristic analysis for intelligent medical systems-a new approach for finding confidence intervals," Biomedical Engineering, IEEE Transactions on, vol. 47, pp. 952-963, 2000.

[5] M. Li and I. K. Sethi, "Confidence-based classifier design," Pattern Recognition, vol. 39, pp. 1230-1240, 2006.

[6] J. Demar, "Statistical comparisons of classifiers over multiple data sets," Journal of Machine Learning Research, vol. 7, pp. 1-30, 2006.

[7] T. Fawcett, "ROC graphs: Notes and practical considerations for researchers," Machine Learning, vol. 31, pp. 1-38, 2004.

[8] S. Fakhraei, H. Soltanian-Zadeh, F. Fotouhi, and K. Elisevich, "Confidence in medical decision making: application in temporal lobe epilepsy data mining," in 2011 workshop on data mining for medicine and healthcare, Co-located with 17th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, San Diego, California, USA, 2011.

[9] T. G. Dietterich, "Ensemble methods in machine learning," in Proceedings of the First International Workshop on Multiple Classifier Systems, June 21-23, 2000, Berlin, Germany, 2000.

[10] S. Fakhraei, H. Soltanian-Zadeh, K. Elisevich, and F. Fotouhi, "Attribute ranking for lateralizing focal epileptogenicity in temporal lobe epilepsy," in 17th Iranian Conference in Biomedical Engineering, ICBME 2010, November 3-4, 2010, Isfahan, Iran, 2010.

[11] K. Jafari-Khouzani, K. Elisevich, S. Patel, and H. Soltanian-Zadeh, "Dataset of Magnetic Resonance Images of Nonepileptic Subjects and Temporal Lobe Epilepsy Patients for Validation of Hippocampal Segmentation Techniques," Neuroinformatics, pp. 1-12, 2011.

[12] K. Jafari-Khouzani, K. Elisevich, S. Patel, B. Smith, and H. Soltanian-Zadeh, "FLAIR signal and texture analysis for lateralizing mesial temporal lobe epilepsy," NeuroImage, vol. 49, pp. 1559-1571, 2010.

[13] K. Jafari-Khouzani, K. Elisevich, K. C. Karvelis, and H. Soltanian-Zadeh, "Quantitative multi-compartmental SPECT image analysis for lateralization of temporal lobe epilepsy," Epilepsy Research, vol. 95, pp. 35-50, 2011.

[14] M. Jenkinson and S. Smith, "A global optimisation method for robust affine registration of brain images," Medical Image Analysis, vol. 5, pp. 143-156, 2001.

[15] S. M. Smith, "Fast robust automated brain extraction," Human Brain Mapping, vol. 17, pp. 143-55, 2002.

[16] S. Fakhraei, H. Soltanian-Zadeh, F. Fotouhi, and K. Elisevich, "Consensus feature ranking in datasets with missing values," in 9th International Conference on Machine Learning and Applications, ICMLA 2010, December 12-14, 2010, Washington, DC, United states, 2010.

[17] A. Arauzo-Azofra, J. L. Aznarte, and J. M. Benitez, "Empirical study of feature selection methods based on individual feature evaluation for classification problems," Expert Systems with Applications, vol. 38, pp. 8170-8177, 2011.

[18] K. Kira and L. A. Rendell, "A practical approach to feature selection," in 9th international workshop on Machine learning, Aberdeen, Scotland, United Kingdom, 1992.