entropy

Article

Noise Robustness Analysis of Performance for EEG-Based Driver Fatigue Detection Using Different Entropy Feature Sets

Jianfeng Hu * and Ping Wang

The Center of Collaboration and Innovation, Jiangxi University of Technology, Nanchang 330098, China; wonderwang1020@jxut.edu.cn

* Correspondence: 200399999@jxut.edu.cn; Tel.: +86-791-88138885

Received: 16 June 2017; Accepted: 25 July 2017; Published: 27 July 2017

Abstract: Driver fatigue is an important factor in traffic accidents, and the development of a detection system for driver fatigue is of great significance. To estimate and prevent driver fatigue, various classifiers based on electroencephalogram (EEG) signals have been developed; however, as EEG signals have inherent non-stationary characteristics, their detection performance is often deteriorated by background noise. To investigate the effects of noise on detection performance, simulated Gaussian noise, spike noise, and electromyogram (EMG) noise were added into a raw EEG signal. Four types of entropies, including sample entropy (SE), fuzzy entropy (FE), approximate entropy (AE), and spectral entropy (PE), were deployed for feature sets. Three base classifiers (K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Decision Tree (DT)) and two ensemble methods (Bootstrap Aggregating (Bagging) and Boosting) were employed and compared. Results showed that: (1) the simulated Gaussian noise and EMG noise had an impact on accuracy, while simulated spike noise did not, which is of great significance for the future application of driver fatigue detection; (2) the influence on noise performance was different based on each classifier, for example, the robust effect of classifier DT was the best and classifier SVM was the weakest; (3) the influence on noise performance was also different with each feature set where the robustness of feature set FE and the combined feature set were the best; and (4) while the Bagging method could not significantly improve performance against noise addition, the Boosting method may significantly improve performance against superimposed Gaussian and EMG noise. The entropy feature extraction method could not only identify driver fatigue, but also effectively resist noise, which is of great significance in future applications of an EEG-based driver fatigue detection system.

Keywords: driver fatigue; electroencephalogram (EEG); bagging; boosting; entropy

1. Introduction

As EEG signals can reflect the instant state of the brain, it is an excellent method to evaluate the state and function of the brain, and is often used to assist in the diagnosis of stroke, epilepsy, and seizure. Various computational methods based on EEG signals have been developed for the analysis and detection of driver fatigue.

Correa et al. [1] developed an automatic method to detect the drowsiness stage in EEG signals using 19 features and a Neural Network classifier, and obtained an accuracy of 83.6% for drowsiness detections. Mu et al. [2] employed fuzzy entropy for feature extraction and an SVM classifier to achieve an average accuracy of 85%. Other results from their study showed that four feature sets (SE, AE, PE, and FE) and SVM were proposed, with an average accuracy of 98.75% [3]. Fu et al. [4] proposed a fatigue detection model based on the Hidden Markov Model (HMM), and achieved a highest accuracy of 92.5% based on EEG signals and other physiological signals. Li et al. [5] collected

16 channels of EEG data and computed 12 types of energy parameters, and achieved a highest accuracy of 91.5%. Xiong et al. [6] combined features of AE and SE with an SVM classifier and achieved a highest accuracy of 91.3%. Chai et al. [7] presented an autoregressive (AR) model for features extraction and a Bayesian neural network for the classification algorithm, and achieved an accuracy of 88.2%. In another study, Chai et al. [8] employed AR modeling and sparse-deep belief networks to yield an accuracy of 90.6%. Chai et al. [9] also explored power spectral density (PSD) as a feature extractor and fuzzy swarm based-artificial neural network (ANN) as a classifier, achieving an accuracy of 78.88%. Wu et al. [10] proposed an online weighted adaptation regularization for a regression algorithm which could significantly improve performance. Huang's [11] results validated the efficacy of this online closed-loop EEG-based fatigue detection.

With respect to driver fatigue detection based on EEG signals, the performance of many linear and nonlinear single classifiers has already been assessed, such as the Fisher discriminant analysis, DT, SVM, KNN, Neural Network, and Hidden Markov Model. However, it may be difficult to build an excellent single classifier as EEG signals are unstable and the training set is usually comparatively small. Consequently, single classifiers may have a poor performance or be unstable. Recent studies have shown that ensemble classifiers perform better than single classifiers [12–15]; however, few studies have been conducted for using ensemble classifiers based on EEG signals to study driver fatigue detection. Hassan and Bhuiyan [12] proposed an EEG based method for sleep staging using Complete Ensemble Empirical Mode Decomposition with Adaptive Noise and Bootstrap Aggregating (Bagging), and their results showed that the proposed method was superior when compared to the state-of-the-art methods in terms of accuracy. Furthermore, Hassan and Subasi [13] implemented linear programming Boosting to perform seizure detection, which performed better than the existing works. Sun et al. [14] evaluated the performance of the three ensemble methods for EEG signal classification of mental imagery tasks with the base classifiers of KNN, DT, and SVM, where their results suggested the feasibilities of ensemble classification methods. Finally, Yang et al. [15] proposed a gradient Boosting decision tree (GBDT) to classify aEEG tracings.

It is well known that EEG signals are non-stationary. The non-stationary signals can be observed during the change in eye blinking, event-related potential (ERP), and evoked potential [16]. Unfortunately, EEG recordings are often contaminated by different forms of noises, such as noises due to electrode displacement, motion, ocular activity, and muscle activity. These offending noises not only misinterpret underlying neural information processing, but may also themselves be difficult to identify [17]. This is one of the major obstacles in EEG signal classification, thus, a classifier optimized for one set of training EEG data may not work with another set of test EEG data. The variety of artifacts and their overlap with signals of interest in both the spectral and temporal domains, and even sometimes in the spatial domain, makes it difficult for a simple signal preprocessing technique to identify them from the EEG. Therefore, the use of simple filtering or amplitude thresholds to remove artifacts often results in poor performance both in terms of signal distortion and artifact removal. Thus, many methods and algorithms have been developed for artifact detection and removal from EEG signals [18,19]; however, some noise removal methods may also weaken features. Our question was to ask if there was a feature extraction method or algorithm that did not need to remove noise, and was insensitive to noise. Thus, this method could improve classification performance, reduce computational complexity and avoid new noise.

Recently, entropy has been broadly applied in the analysis of EEG signals as EEG is a complex, unstable, and non-linear signal [20,21]. A diverse collection of these methods has been proposed in the last few decades, including spectral entropy (PE), permutation entropy, distribution entropy, fuzzy entropy (FE), Renyi entropy, approximate entropy (AE), sample entropy (SE) and others. Specifically in the field of EEG processing, four of the most widely used and successful entropy estimators are FE [22], AE [23], and SE [24]. AE has demonstrated its capability to detect complex changes; SE is a similar statistic, but has not yet been used as extensively as AE. AE and SE are very successful data entropy estimators, but they also have their weaknesses. AE is biased since it includes self-matches in the count, and SE requires a relatively large r to find similar subsequences and to avoid the log(0) problem. They are also very sensitive to input parameters m, r, and N [25]. More recently, FE has been proposed to alleviate these problems. FE is based on a continuous function to compute the dissimilarity between two zero-mean subsequences and, consequently, is more stable in noise and parameter initialization terms. These metrics is still scarcely used in EEG studies, but are expected to replace AE and SE because of their excellent stability, mainly when applied to noisy or short records.

Given the non-stationary characteristics of EEG signals, we have observed that the optimal detection performance varied as a result of the classifiers or feature sets, which is a major obstacle in EEG signal classification. Thus, a classifier optimized for a particular set of training data may not work well for driver fatigue detection with new data.

Investigating the ability of feature sets and classifiers to evaluate the performance of a detection system in the presence of noise is an important area of investigation as the real EEG signal is seldom noise free. However, how the addition of simulated noise can cause changes in the driver fatigue detection performance for various classifiers or various feature sets has yet to be sufficiently studied. Furthermore, research involving noise robustness analysis to evaluate for the driver fatigue detection performance of the EEG signals in the presence of noise by various feature sets and various classifiers has not been addressed. In general, systematic study investigating the effects of simulated noise on driver fatigue detection systems and the ability of such measures to evaluate the detection systems under simulated Gaussian noise is missing. To the best of our knowledge, our study is one of the first to apply the noise robustness analysis method on EEG signals for driver fatigue detection.

In this study, our aim was to evaluate the robustness of various classifiers and feature sets for driver fatigue detection systems under simulated Gaussian noise. Four types of entropy were deployed as feature sets in this work: FE, SE, AE, and PE. The classification procedure was implemented by three base classifiers: KNN, SVM, and DT, which have been known as state-of-the-art classification methods in many studies. The ensemble classifiers were developed by two ensemble methods: Bagging and Boosting. The challenge was to analyze the impacts of noise on detection performance with four feature sets and five classification methods.

First, with simulated Gaussian noise, we compared the detection performance, i.e., the average accuracy of DT, SVM, and KNN methods. Second, we evaluated the noise robustness of these methods. The noisy EEG signals were generated with the addition of random Gaussian noise into the original EEG signal. Then, we assessed the noise robustness of these methods. Third, in addition to the base classifiers, we examined the effects of the Bagging and Boosting ensemble methods. Moreover, we repeated these analyses with simulated spike noise and simulated EMG noise. This paper is organized as follows: in Section 2, the experiment and EEG signal processing methods such as acquisition, preprocessing, segment, feature extraction and classification are described. In addition, noise generation is explained in this section. Section 3 shows the experimental results and discussion. Finally, we conclude this paper in Section 4.

2. Materials and Methods

Figure 1 shows the workflow of this paper, including EEG acquisition, preprocessing, segment, feature extraction, noise generation, classification, and performance analysis.

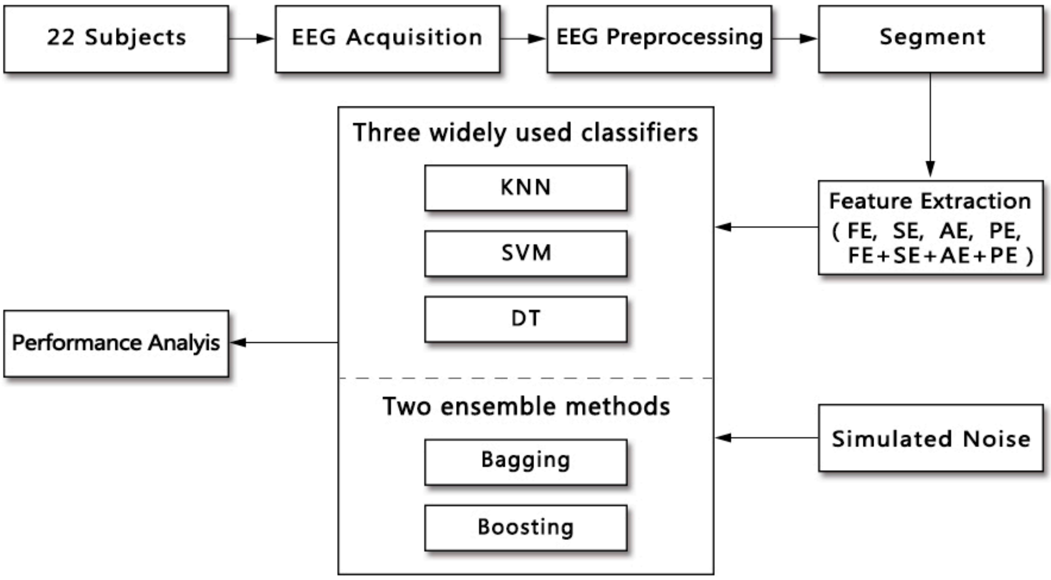


Figure 1. Workflow of proposed study.

2.1. Subjects

Twenty-two university students (14 male, 19–24 years) participated in this experiment. All subjects were asked to be abstain from any type of stimulus like alcohol, medicine, or tea before and during the experiment. Before the experiment, subjects practiced the driving task for several min to become acquainted with the experimental procedures and purposes. This work was approved by all subjects, and the experiments was authorized by the Academic Ethics Committee of the Jiangxi University of Technology. The subjects provided their written informed consent as per human research protocol in this study. Furthermore, all subjects provided their written informed consent as per human research protocol in this study.

2.2. Experimental Paradigm

The driving fatigue simulation experiment was performed by each subject on a static driving simulator (The ZY-31D car driving simulator, produced by Peking ZhongYu CO., LTD, Beijing, China), as shown in Figure 2. On the screen, a customized version of the Peking ZIGUANGJIYE software ZG-601 (Car driving simulation teaching system, V9.2) was shown.

This equipment was an analog form of a real driving car, which contained all the driving capabilities of a vehicle. Using computer software technology, different driving environments could be constructed, such as sunny, foggy or snowy weather and mountain, highway, and countryside areas. The driving environment selected for this experiment was a highway with low traffic density that could more easily induce monotonous driving. Some research has suggested that the brain in this driving environment is more easily turned into a state of fatigue and the EEG signal was more stable, therefore benefiting our next data recording. All subjects in this experiment had an approximate real driving experience.

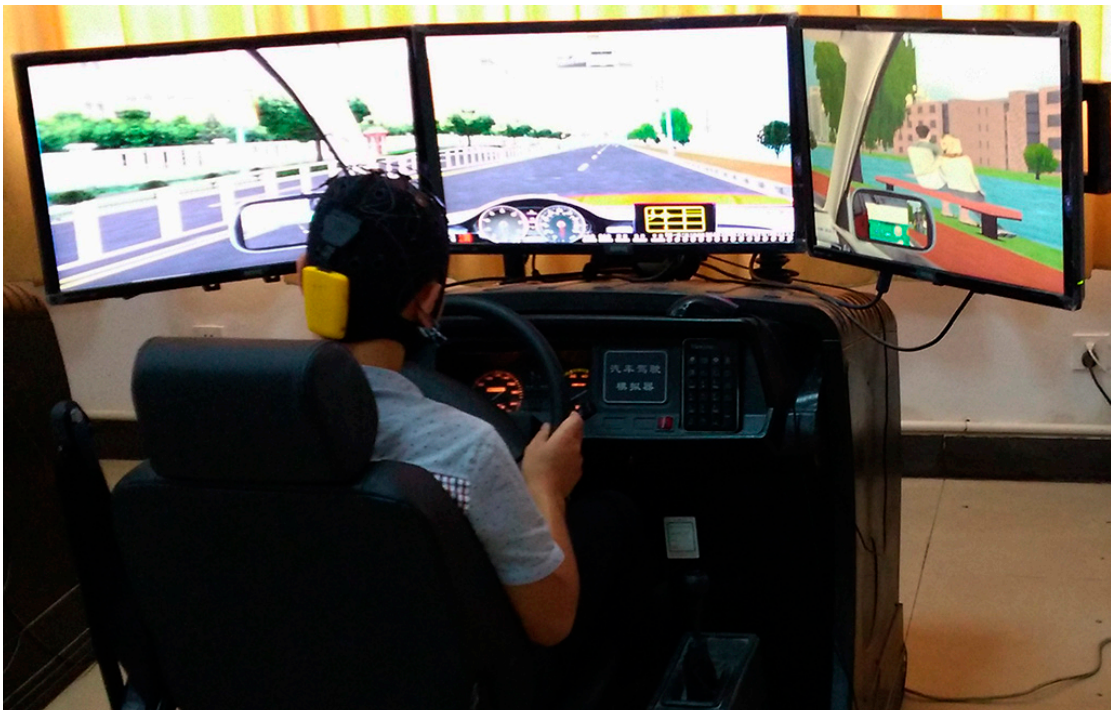


Figure 2. Snapshot of the experimental setup.

2.3. Data Acquisition and Preprocessing

In summary, the total duration of the experiment was 40–130 min. The first step was to become familiar with the simulating software, followed by continuous monotonous driving until driver fatigue was determined and the experiment terminated.

When the driving lasted 10 min, the last 5 min of the EEG signals were recorded as the normal state. When the continuous driving lasted 30–120 min (until the self-reported fatigue questionnaire results showed the subject was in driving fatigue), obeying Borg's fatigue scale and Lee's subjective fatigue scale, the last 5 min of the EEG signals were labeled as the fatigue state. EOG was also used to analyze eye blink patterns as an objective part of the validation of the fatigue state. It should be noted that the validation of the fatigue condition was also based on a self-reported fatigue questionnaire as per Borg's fatigue scale and Lee's subjective fatigue scale [26,27]. This method of using a questionnaire to identify the fatigue condition has not only been used in our study, but also in many other studies [2,3]. The drivers were required to complete all tasks and ensure safe driving. Prior to the experiment, the drivers familiarized themselves with the driving simulator and the completion of the driving tasks.

All channel data were referenced to two electrically linked mastoids at A1 and A2, digitized at 1000 Hz from a 32-channel electrode cap (including 30 effective channels and two reference channels) based on the International 10–20 system (Figure 3) and stored in a computer for offline analysis. Eye movements and blinks were monitored by recording the horizontal and vertical EOG.

After the acquisition of EEG signals, the main steps of data preprocessing were carried out by using the Scan 4.3 software of Neuroscan (Compumedics, Australia). The raw signals were first filtered by a 50 Hz notch filter and a 0.15–45 Hz band-pass filter was used. Next, 5-min EEG signals from 30 channels were sectioned into 1-s epochs, resulting in 300 epochs. With 22 subjects, a total of 6600 epochs (792,000 units for 30 channels and 4 feature sets) of dataset was randomly formed for the normal state and another 6600 epochs (792,000 units for 30 channels and 4 feature sets) for the fatigue state.

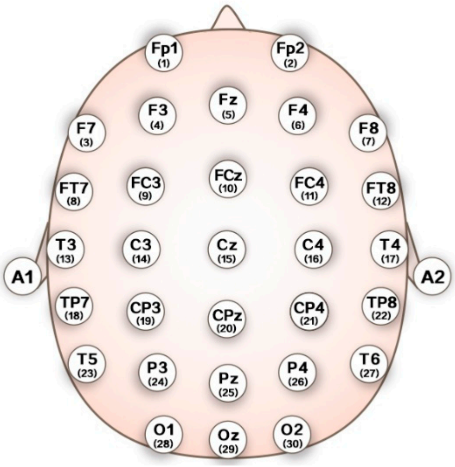


Figure 3. Electrodes position as per International 10–20 System standard.

2.4. Feature Extraction

As the EEG signal is assumed to be a non-stationary time series and most feature extraction methods are only applicable to stationary signal, in this study, to deal with this problem, the EEG time series was divided into many short windows and its statistics were assumed to be approximately stationary within each window. The following feature extraction methods were applied to each 1 s windowed signal. EEG signals were segmented without overlap, and finally feature sets were extracted from all channels in each 1 s window.

The ability to distinguish between the normal state and fatigue state depended mainly on the quality of input vectors of the classifier. To capture EEG characteristics, four feature sets including FE, SE, AE, and PE were calculated [21–25]. In this section, methods for the calculation of these feature sets on EEG recordings are described in detailed.

2.4.1. Spectral Entropy (PE)

PE was evaluated using the normalized Shannon entropy [28], which quantifies the spectral complexity of the time series. The power level of the frequency component is denoted by Y_i and the normalization of the power y_i is performed as:

$$y_i = \frac{Y_i}{\sum Y_i} \tag{1}$$

The spectral entropy of the time series is computed using the following formula:

$$\mathrm{PE} = \sum_i y_i log(\frac{1}{y_i}) \tag{2}$$

2.4.2. Approximate Entropy (AE)

AE, as proposed by Pincus [23], is a statistically quantified nonlinear dynamic parameter that measures the complexity of a time series. The procedure for the AE-based algorithm is described as follows:

Considering a time series t(i), a set of m-dimensional vectors are obtained as per the sequence order of t(i):

$$T_i^m = [t(i),\ t(i+1),\ \cdots\ ,t(i+m-1)];\ 1 \le L-m+1 \tag{3}$$

where $d[T_i^m, T_j^m]$ is the distance between two vectors T_i^m and T_j^m, defined as the maximum difference values between the corresponding elements of two vectors:

$$d[T_i^m, T_j^m] = \max_{k\in(0,m-1)}\{|t(i+k) - t(j+k)|\}, (i, j = 1 \sim L - m + 1, i \neq j) \tag{4}$$

Define S_i as the number of vectors T_j that are similar to T_i, subject to the criterion of similarity $d[T_i^m, T_j^m] \leq s$

$$S_i^m(s) = \frac{1}{L - m + 1} S_i \tag{5}$$

Define the function $\gamma^m(s)$ as:

$$\gamma^m(s) = \frac{1}{L - m + 1} \sum_{i=1}^{L-m+1} \ln S_i^m(s) \tag{6}$$

Set m = m + 1, and repeat Equations (1) to (3) to obtain $S_i^{m+1}(s)$ and $\gamma^{m+1}(s)$, then:

$$\gamma^{m+1}(s) = \frac{1}{L - m} \sum_{i=1}^{L-m} lnS_i^{m+1}(s) \tag{7}$$

The approximate entropy can be expressed as:

$$AE = \gamma^m(s) - \gamma^{m+1}(s) \tag{8}$$

2.4.3. Sample Entropy (SE)

The SE algorithm is like that of AE [25,29], and is a new measure of time series complexity proposed by Richman and Moorman [24]. Equations (1) and (2) can be defined in the same way as the AE-based algorithm; other steps in the SE-based algorithm are described as follows:

Define A_i as the number of vectors T_j that are similar to T_i, subject to the criterion of similarity $d[T_i^m, T_j^m] \leq s$

$$A_i^m(s) = \frac{1}{L - m - 1} A_i \tag{9}$$

Define the function $\gamma^m(s)$ as:

$$\gamma^m(s) = \frac{1}{L - m} \sum_{i=1}^{L-m} A_i^m(s) \tag{10}$$

Set m = m + 1, and repeat the above steps to obtain $A_i^{m+1}(s)$ and $\gamma^{m+1}(s)$, then

$$\gamma^{m+1}(s) = \frac{1}{L - m} \sum_{i=1}^{L-m} A_i^{m+1}(s) \tag{11}$$

The sample entropy can be expressed as:

$$SE = log(\gamma^m(s)/\gamma^{m+1}(s)) \tag{12}$$

2.4.4. Fuzzy Entropy (FE)

To deal with some of the issues with sample entropy, Xiang et al. [22] proposed the use of a fuzzy membership function in computing the vector similarity to replace the binary function in sample entropy algorithm, so that the entropy value as continuous and smooth. The procedure for the FE-based algorithm is described in detail as follows:

Set a L-point sample sequence: $\{v(i) : 1 \leq i \leq L\}$;

The phase-space reconstruction is performed on $v(i)$ as per the sequence order. The reconstructed vector can be written as:

$$T_i^m = \{v(i), v(i+1), \ldots, v(i+m-1)\} - v_0(i) \tag{13}$$

where $i = 1, 2, \ldots, L-m+1$, and $v_0(i)$ is the average value described as the following equation:

$$v_0(i) = \frac{1}{m}\sum_{j=0}^{m-1} v(i+j) \tag{14}$$

d_{ij}^m, the distance between two vectors T_i^m and T_j^m, is defined as the maximum difference in values between the corresponding elements of two vectors:

$$\begin{gathered} d_{ij}^m = \mathrm{d}[T_i^m, T_j^m] = \max_{k\in(0,m-1)}\{|v(i+k) - v_0(i) - (v(j+k) - v_0(j))|\} \\ (i, j = 1 \sim L-m, i \neq j) \end{gathered} \tag{15}$$

Based on the fuzzy membership function $\sigma(d_{ij}^m, n, s)$, the similarity degree D_{ij}^m between two vectors T_i^m and T_j^m is defined as:

$$D_{ij}^m = \sigma(d_{ij}^m, n, s) = \exp(-(d_{ij}^m)^n / s) \tag{16}$$

where the fuzzy membership function $\sigma(d_{ij}^m, n, s)$ is an exponential function, while n and s are the gradient and width of the exponential function, respectively.

Define the function $\gamma^m(n, s)$:

$$\gamma^m(n,s) = \frac{1}{L-m}\sum_{i=1}^{L-m}\frac{1}{L-m-1}\sum_{j=1, j\neq 1}^{L-m} D_{ij}^m \tag{17}$$

Repeat the Equations (1) to (4) in the same manner. Define the function:

$$\gamma^{m+1}(n,s) = \frac{1}{L-m}\sum_{i=1}^{L-m}\frac{1}{L-m-1}\sum_{j=1, j\neq 1}^{L-m} D_{ij}^{m+1} \tag{18}$$

The fuzzy entropy can be expressed as:

$$FE(m, s, n) = ln\gamma^m(n, s) - ln\gamma^{m+1}(n, s) \tag{19}$$

In the above-mentioned four types of entropies, AE, SE and FE have variable parameters, m and r. In the present study, m = 2 while r = 0.2*SD, where SD denotes the standard deviation of the time series as per the literature [3,22,25].

For optimizing detection quality, the feature sets were normalized for each subject and each channel by scaling between 0 and 1.

2.5. Classification

However, due to the lack of a substantial sample size, algorithms based on ensemble learning methods needed to evaluate the detection performance for driver fatigue. Bagging is an acronym of "bootstrap aggregating" [30,31], and builds several subsets and aggregates their individual predictions to form a final prediction. In the Bagging method, the number of base classifiers must be set. To investigate the impact of base classifier number on the classification result, we set the number of base classifiers as 50, 100, and 200, respectively. Like Bagging, Boosting also uses subsets to train classifiers, but not randomly [32–34]. In Boosting, difficult samples have higher probabilities of being selected for training, and easier samples have less chance of being used. In the Boosting method, the number of Boosting stages has to be set. To investigate the impact of the Boosting stage number on the classification result, we set the number of the Boosting stage to 50, 100, and 200, respectively.

The Bagging and Boosting methods both try to construct multiple classifiers by using different subsets. Bagging trains each classifier over a randomly selected subset, while the Boosting method trains each new classifier [35].

Some classification models can fit data for a range of values of a parameter almost as efficiently as fitting the classifier for specific value of the parameters. This feature can be leveraged to perform a more efficient cross-validation for the selection of parameters. A high variance can lead to over-fitting in model selection, and hence poor performance, even when the number of hyper-parameters is relatively small [36]. It seems likely that over-fitting during model selection can be overcome using various approaches. To overcome the bias in performance evaluation, parameter selection should be conducted independently in each trial to prevent selection bias and to reflect optimal performance. Performance evaluation based on these principles requires repeated training with different sets of hyper-parameter values on different samples of the available data, which makes it well-suited to parallel implementation. The magnitude of the bias deviations from full nested cross-validation can be introduced, which can easily swamp the difference in performance between the classifier systems.

To avoid the problem of over-fitting and to make general classifiers for other independent datasets, the datasets were separated into training sets and test sets in the following pattern. In the training phase, a 10-fold cross validation was applied on the features so that 10% of the feature vectors were dedicated as a test set and the other 90% of feature vectors were considered as the training set. In the next iteration, another 10% of the feature vectors were considered as a test set and the rest for the training set, until all the feature vectors had participated once in the test phase. The final result was achieved by averaging the outcome produced in the corresponding test repeated 10 times (for different subjects and different feature sets). Using this evaluation scheme, the dependency of the training and test features was removed, thus avoiding the over-fitting problem [37–42]. In particular, though GB is a more capable and practical boosting algorithm, like most other classifiers, GB also had the problem of over-fitting when dealing with very noisy data. To overcome such a problem, the validation sets were used to adjust the hypothesis of the Boost algorithm to improve generalization, thereby alleviating overfitting and improving performance, which have long been used in addressing the problem of overfitting with neural networks and decision trees [43,44]. Its basic concept is to apply the classifier to a set of instances distinct from the training set. Thus, the sequence of base classifiers produced by GB from the training set, also is applied to the validation set for alleviating overfitting problem.

For optimizing parameters, it is very important to obtain the optimum values for the classifier performance. Three widely used classifiers (KNN, SVM, and DT) were employed as classifiers in this work. To select optimal parameters of the model, this paper adopted the method of cross validation based on grid search, thus avoiding arbitrary and capricious behavior. Grid search is a model hyperparameter optimization technique. In this study, a grid parameter search was used to achieve optimal results. Related parameters in this study are as follows: penalty parameter, kernel and kernel coefficient for SVM, number of neighbors for KNN, the number of features, the maximum depth of the tree and the minimum number of samples for DT, the number of base estimators and the number of features for Bagging method, learning rate, the number of boosting stages and maximum depth for Boosting method.

2.6. Simulated Noise

The noises of the EEG signals included white noise, spike noise, muscular noise, ocular noise, and cardiac noise.

2.6.1. White Noise

White noise accounts for possible sources in real environments, such as thermal noise or electro-magnetic noise, which can be generated by a Gaussian random process. Spikes can be of sensor movement origin and the probability of appearance was kept relatively low in a real case. Muscular artifacts were drawn from electromyogram (EMG) signals. Ocular artifacts came from electrooculogram (EOG) signals. Cardiac artifacts were generated by heartbeat. In this paper, only white noise was considered for simplicity.

To analyze the influence of noise on detection performance, we built a simulated Gaussian noise P_i^{noise}:

$$P_i'(t) = P_i(t) + P_i^{noise} \tag{20}$$

where P_i is the original EEG signal of channel I; P_i^{noise} is the simulated Gaussian white noise; and P_i' is the noisy EEG signal with simulated Gaussian white noise. We assumed that P_i^{noise} and P_i were uncorrelated, and P_i^{noise}~D*N(0, 1). Here, D is defined as the level of noise given as a percentage of the average level of the noise-free data $P_i(t)$.

Therefore, to evaluate the noise robustness of the classifiers systematically, we used scale factor D to control the noise power. To make polluted EEG data by Gaussian noise, we generated the same dimension of Gaussian noise to the segmented EEG signal, i.e., noise dimension was 1024 per second per channel.

2.6.2. Spike Noise

Spikes were synthetically generated as described in Reference [45] and these interferences can be of a technological (sensor movement, electrical interferences) or physiological (mainly eye blinks) origin. The probability of appearance was kept relatively low (0.01), as to be expected in a real case. Duration was set at 1 sample and amplitude was set at 1.

2.6.3. Muscular Noise

Muscular noises were drawn from an actual long electromyogram (EMG) signal downloaded from PhysioNet [46], which corresponded to a patient with myopathy. Data were acquired at 50 KHz and then down sampled to 1 KHz. For each run, an EMG epoch of length N was extracted from the entire record by commencing at a random sample. These noises accounted for muscular activity during EEG recording.

2.7. Performance Metrics

To estimate the potential application performance of a detector, it is very important to properly examine the detection quality. The total average accuracy based on a feature set and some classifiers was the average of the accuracy of all single channels based on the same feature and the same classifiers. The classification capabilities of different classifiers were comprehensively investigated with several indexes including Accuracy, Precision, Recall, F1-score, and the Matthews Correlation Coefficient (MCC) [47]. These indexes are given as follows: Accuracy is the percentage of normal predictions corresponding to all samples; Precision is the percentage of normal predictions corresponding to the normal samples; and Recall is the percentage of fatigue predictions corresponding to the fatigue samples. Furthermore, the F1-score was used to appraise both Precision and Recall. The MCC was used as a measure of the quality of binary classifications as it considers true and false positives and negatives, and is generally regarded as a balanced measure which can be used even if the classes are of extremely different sizes. Therefore, a high Precision, Recall, F1-score, and MCC value relates to higher performance. The following equation set is used in the literature for examining performance quality:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{21}$$

The recall is intuitively the ability of the classifier to find all the positive samples.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{22}$$

The precision is intuitively the ability of the classifier not to label as positive a sample that is negative.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{23}$$

The F1 score can be interpreted as a weighted average of the precision and recall, where an F1 score reaches its best value at 1 and worst score at 0.

$$\text{F1} - \text{score} = \frac{2TP}{2TP + FN + FP} \tag{24}$$

The Matthews correlation coefficient (MCC) is used in machine learning as a measure of the quality of two-class classifications. The MCC is a correlation coefficient value between −1 and +1.

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{25}$$

where *TP* (true positive) represents the number of normal signals identified as such; *TN* (true negative), the number of fatigue signals classified as such; *FP* (false positive), the number of fatigue signals recognized as such; *FN* (false negative), the number of normal signals distinguished as fatigue signals.

To investigate differences in average accuracy among various classifiers and feature sets, the paired sample *t*-test was used to evaluate the effectiveness of each comparison. The results were averaged over ten independently drawn combinations in each experiment.

3. Results and Discussion

3.1. Gaussian Noise

In general, when all EEG channels are used for detecting driver fatigue, good results may be achieved; however, we wanted to understand what the impact would be on detection performance if the noise was superimposed on some channels. To investigate this question, first, the feature sets (FE, SE, AE and PE) of the EEG signals across all 30 channels were extracted for training and recognition, before gradually adding a number of noisy channels.

To explore the effect of the number of noisy channels that can be added to the detection system, we evaluated the system performance with respect to the number of polluted channels. For each number m (from 1 to 30), a random combination (m out of 30 channels) was repeated 10 times to calculate classification accuracy using a 10-fold cross validation. The scale factor (D) of superimposed noise was set at 1.0. Furthermore, for each condition (m from 1 to 30), the paired *t*-test was used as a post-hoc test to evaluate and compare the performance of the classifiers.

3.1.1. Effect of Noise: DT Classifier vs. KNN Classifier vs. SVM Classifier

Based on the literature [36–39], of the four feature sets, FE out-performed the other feature sets. DT was the best among several classifiers, while SVM was the weakest. Here, we compare the detection performance for the three classifiers and four feature sets with increasingly noisy channels. The comparison among the three classifiers in terms of average accuracy for each feature set is shown in Figure 4.

First, we evaluated the classification accuracy of these methods using the original experimental datasets uncontaminated by noise sources. We observed little difference in the average accuracy between the three classifiers; moreover, we investigated the impact of increasing the noisy channels on the detection performance of each method. The number of noisy channels varied from 1 to 30.

When using the FE feature set, there were no differences in average accuracy among KNN, DT, and SVM for the original EEG signals (paired t-test, $p > 0.05$). However, with more channels adding noise, the average accuracy of the three classifiers decreased, but DT decreased slowly. For the DT classifier, the average accuracy was decreased from 0.958 with a noise-free signal to 0.771 with 30 noisy channels. For the SVM and KNN, the average accuracies were 0.972 and 0.966 with noise-free signals and dropped to 0.682 and 0.590 with 30 noisy channels, respectively. The performance of DT was better than those of SVM and KNN in the presence of 30 noisy channels. In addition, the mean difference of the classification accuracy between DT and SVM (KNN) was statistically significant in the presence of 30 noisy channels (paired t-test, $p < 0.01$).

When using feature set AE, SVM achieved a competitive average accuracy over both DT and KNN for the original EEG signals. However, this difference was not statistically significant (paired t-test, $p > 0.05$). After noise addition by the proposed method, significantly lower accuracy was obtained by the SVM than by DT. However, with more and more channels adding noise, the average accuracy of three classifiers decreased, but DT decreased slowly. For the DT classifier, the average accuracy was decreased from 0.929 with noise-free signal to 0.690 with 30 noisy channels. For the SVM and KNN, the average accuracies were 0.952 and 0.926 with noise-free signal and dropped to 0.647 and 0.550 with 30 noisy channels, respectively. The effect of DT was better than those of SVM and KNN in the presence of 30 noisy channels (paired t-test, $p < 0.01$). The effect of SE was similar to that of AE.

When using the feature set PE, the detection performance of SVM was significantly better than the other two classifiers for the original EEG signals (paired t-test, $p < 0.01$). However, with more and more channels adding noise, the average accuracy of three classifiers decreased, until the final average accuracy was almost the same. For the DT classifier, the average accuracy decreased from 0.782 with a noise-free signal to 0.636 with 30 noisy channels (paired t-test, $p < 0.01$). For the SVM and KNN, the average accuracies were 0.825 and 0.763 with noise-free signal and dropped to 0.645 and 0.567 with 30 noisy channels, respectively (paired t-test, $p < 0.01$).

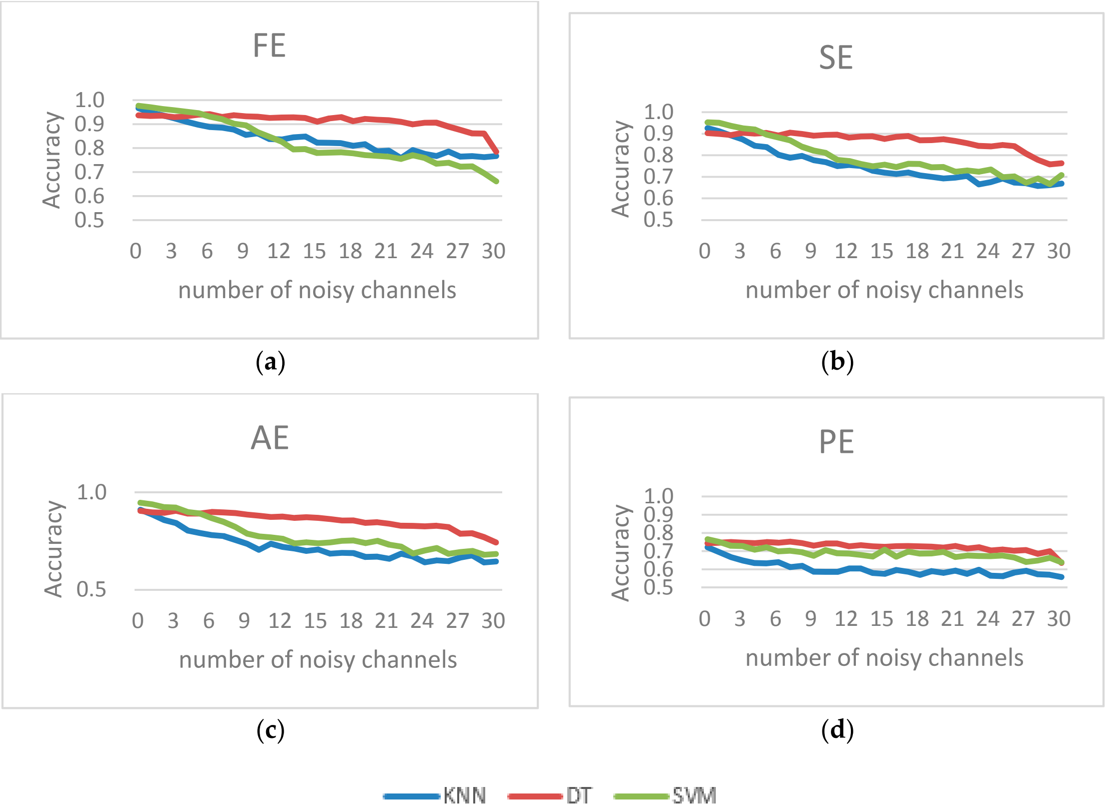
From the above results, for all four feature sets (FE, SE, AE and PE), the difference between the various classifiers for noisy EEG signals was clear and remained consistent, with the performance of the DT classifier being greater than those of the SVM and KNN classifiers (except PE). The average accuracy for the DT classifier decreased slowly, while the average accuracy for the other two classifiers decreased quickly with increasing noisy channels. As mentioned before, the difference between the DT classifier and the other two classifiers continued to grow across a varying number of noisy channels, with little difference in the classification accuracy between the SVM and KNN methods.

The classification accuracy of the FE consistently out-performed other feature sets regardless of classifier. The most significant differences in the noisy EEG data between SVM and DT were found in FE. The average accuracy for the PE feature set with original noise-free EEG was not high, and was lower than those for the other three feature sets significantly (paired t-test, $p < 0.01$). In two-class classification problems, the theoretical chance level is 50%; however, in the EEG based driver fatigue detecting system, classification accuracy of at least 60% is considered as a threshold for an acceptable recognition. Thus, there is little difference among the three classifiers with the PE feature set.

3.1.2. Effect of Noise: Using Bagging Ensemble Learning Method

As mentioned before, many studies have found that the use of ensemble learning can provide a certain degree of robustness for noise; nevertheless, we wanted to investigate whether ensemble learning would work for driving fatigue detection. Next, we analyzed the effect of noise using the Bagging ensemble learning method. A comparison between average accuracies obtained from noisy EEG data using the Bagging method is illustrated in Figure 5.

As shown in Figure 5, bKNN50, bKNN100 and bKNN200 represent the Bagging ensemble with 50, 100, and 200 KNN base classifiers, respectively. Figure 5 shows how the average accuracy for different classifiers and feature sets changed with an increase in noisy channels. The average accuracy for the Bagging method decreased the same as that for KNN without the Bagging method when noisy channels increased. There was no difference in average accuracy between the KNN without the Bagging method and KNN with the Bagging method (paired *t*-test, $p > 0.05$), and average accuracies both decreased with the increase in noisy channels.



(**a**) (**b**)

(**c**) (**d**)

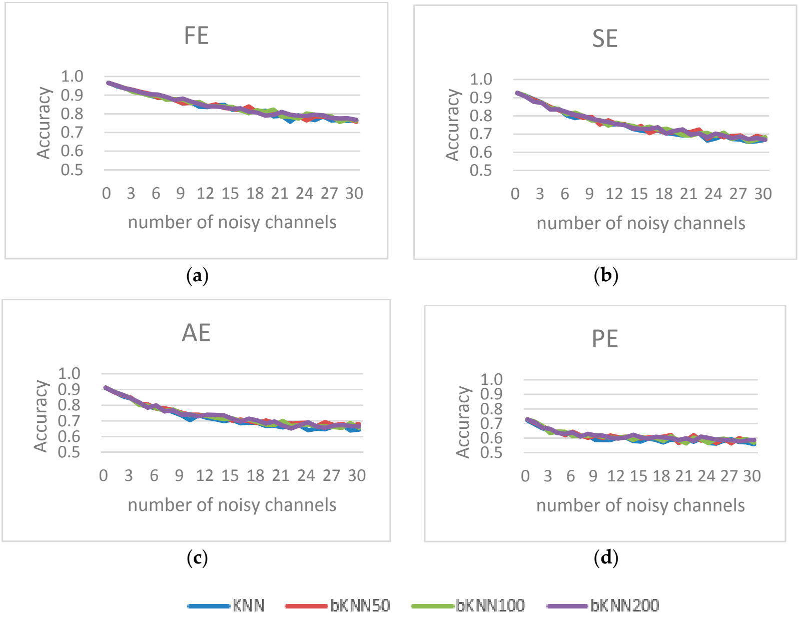
Figure 4. Influence of superimposed noise on the average accuracy for three classifiers when using (**a**) FE feature set; (**b**) SE feature set; (**c**) AE feature set and (**d**) PE feature set. The left vertical coordinate is for average accuracy, while the horizontal coordinate is for number of noisy channels.

When using the feature set FE, there was no difference in average accuracy between KNN without the Bagging method and KNN with the Bagging method (paired *t*-test, $p > 0.05$). However, with more channels adding noise, the average accuracy of KNN classifiers without the Bagging method and with the Bagging method both decreased. For the KNN classifier, the average accuracy decreased from 0.966 with a noise-free signal to 0.590 with 30 noisy channels. For the Bagging method with 50, 100, and 200 base classifiers, the average accuracies were 0.954, 0.954, and 0.954 with noise-free signal and dropped to 0.534, 0.561 and 0.550 with 30 noisy channels, respectively. The other three feature sets were similar to FE.

The above results show that the Bagging method cannot effectively improve the recognition effects of the KNN classifier without noisy channels and with noisy channels. There was also no obvious effect when the number of base classifiers was increased.

3.1.3. Effect of Noise: Using Boosting Ensemble Learning Method

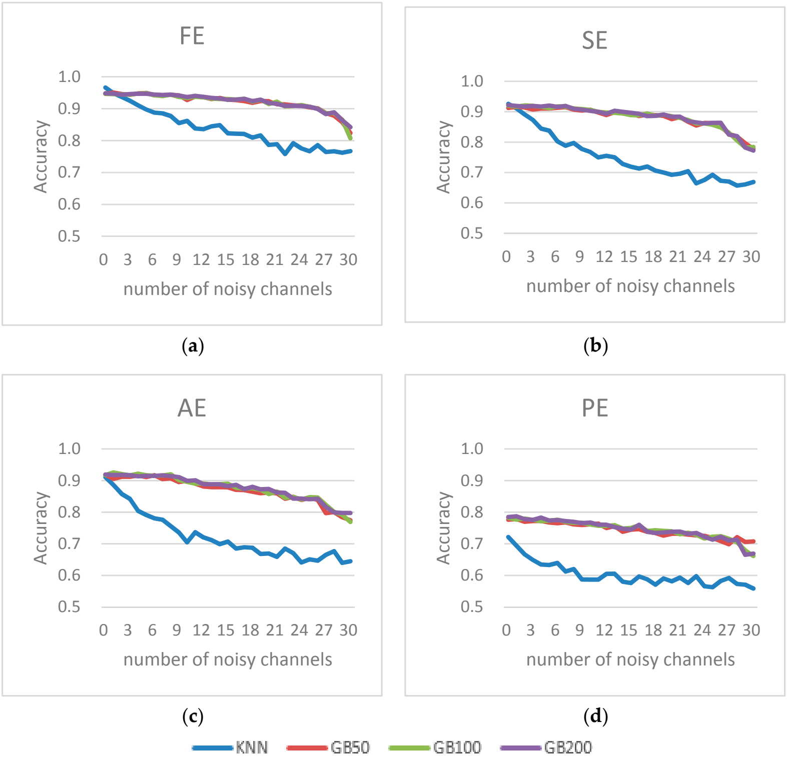
Next, we analyzed the effects of noise using another ensemble learning method, Boosting. As shown in Figure 6, GB50, GB100 and GB 200 represent the Boosting ensemble with 50, 100, and 200 Boosting stages.



(**a**) (**b**)

(**c**) (**d**)

Figure 5. Influence of added noise on the average accuracy for three classifiers with the Bagging method when using (**a**) FE feature set; (**b**) SE feature set; (**c**) AE feature set and (**d**) PE feature set. The left vertical coordinate is for average accuracy, while the horizontal coordinate is for number of noisy channels. bKNN50, bKNN100 and bKNN200 represent the Bagging ensemble method with 50, 100, and 200 K-Nearest Neighbors (KNN) base classifiers.



(**a**) (**b**)

(**c**) (**d**)

Figure 6. Influence of added noise on the average accuracy for three classifiers using the Boosting method when using (**a**) FE feature set; (**b**) SE feature set; (**c**) AE feature set and (**d**) PE feature set. The left vertical coordinate is for average accuracy, while the horizontal coordinate is for the number of noisy channels. GB50, GB100, and GB200 represent the Boosting ensemble with 50, 100, and 200 Boosting stages.

Figure 6 shows how the average accuracy for different classifiers and feature sets changed with increasing channels of additive noise. In the case of the Boosting method, for the four feature sets (FE, SE, AE, and PE), the difference between the various classifiers for noise-free EEG signals and noisy EEG signals was clear (paired *t*-test, $p < 0.01$). The average accuracy for the Boosting method decreased slower than that of KNN without the Boosting method when the noisy channels increased.

When using the feature set FE, there were differences in average accuracy between KNN and the Boosting method (paired *t*-test, $p < 0.01$). With more channels adding noise, the average accuracy of both classifiers decreased, but the Boosting method decreased slowly. For the KNN classifier, the average accuracy decreased from 0.966 with noise-free signals to 0.590 with 30 noisy channels. For the Boosting method with 50, 100, and 200 base classifiers, the average accuracies were 0.950, 0.947 and 0.947 with noise-free signal and dropped to 0.793, 0.806 and 0.792 with 30 noisy channels, respectively. The other three feature sets were similar to FE.

The above results show that the Boosting method was unable to improve the recognition effect without noisy channels; however, it did significantly improve the recognition effect with noisy channels, and there was no obvious effect when the number of base classifiers was increased.

The above results are summarized in Table 1. A_0 is defined as the average accuracy with noise-free signals while A_{30} is defined as the average accuracy with 30 noisy channels signals. In this paper, A_{30}/A_0 is used as an important indicator for robustness. Table 1 summarizes the average accuracy of the three classifiers and the two ensemble methods in the four feature sets obtained from noisy EEG data. It was noted that the Boosting method had significantly different average accuracies from other methods across all feature sets when the EEG data were polluted. Furthermore, FE achieved a better performance than those of SE and AE.

Table 1. Results of the analysis of the average accuracy with simulated noise electroencephalogram (EEG) signals in a 30-channel system. A_0 is defined as the average accuracy with noise-free signals while A_{30} is defined as the average accuracy with 30 noisy channels signals.

A_{30}/A_0	FE	SE	AE	PE
KNN	0.793	0.723	0.708	0.774
DT	0.838	0.845	0.821	0.855
SVM	0.677	0.743	0.722	0.835
bKNN50	0.784	0.730	0.744	0.798
bKNN100	0.791	0.737	0.719	0.781
bKNN200	0.796	0.722	0.729	0.806
GB50	0.869	0.853	0.844	0.912
GB100	0.854	0.853	0.838	0.845
GB200	0.889	0.837	0.867	0.852

3.1.4. Combined Entropy as Feature Sets

Combined entropy has been employed to achieve a better performance [3], but questions remain as to the impact on the detection performance if noise was superimposed on some channels. Combined feature sets (FE + SE + AE + PE) of EEG signals were extracted for training and recognition, before gradually adding noise.

As shown in Figure 7, there were no differences in average accuracy among the KNN, DT, and SVM for the original EEG signals (paired *t*-test, $p > 0.05$); however, with more channels adding noise, the average accuracy of the three classifiers decreased, but DT decreased slower than the others. For the DT classifier, the average accuracy decreased from 0.933 with a noise-free signal to 0.815 with 30 noisy channels. For the SVM and KNN classifiers, the average accuracies were 0.929 and 0.941 with noise-free signals and dropped to 0.688 and 0.762 with 30 noisy channels, respectively. The effect of the DT was better than those of the SVM and KNN in the presence of 30 noisy channels (paired *t*-test, $p < 0.01$). For the Bagging method with 50, 100, and 200 base classifiers, the average accuracies were 0.943, 0.943, and 0.943 with noise-free signals and dropped to 0.778, 0.790, and 0.772

with 30 noisy channels, respectively. The above results show that the Bagging method cannot effectively improve the recognition effect of the KNN classifier without noisy channels and with noisy channels (paired *t*-test, $p > 0.05$). There was also no obvious effect when the number of base classifiers was increased. For the Boosting method with 50, 100, and 200 base classifiers, the average accuracies were 0.944, 0.942, and 0.953 with noise-free signals and dropped to 0.900, 0.888 and 0.878 with 30 noisy channels, respectively. The above results show that the Boosting method was unable to improve the recognition effect without noisy channels (paired *t*-test, $p > 0.05$); however, it could significantly improve the recognition effect with noisy channels (paired *t*-test, $p < 0.01$). Furthermore, there were no obvious effects when the number of base classifiers was increased.

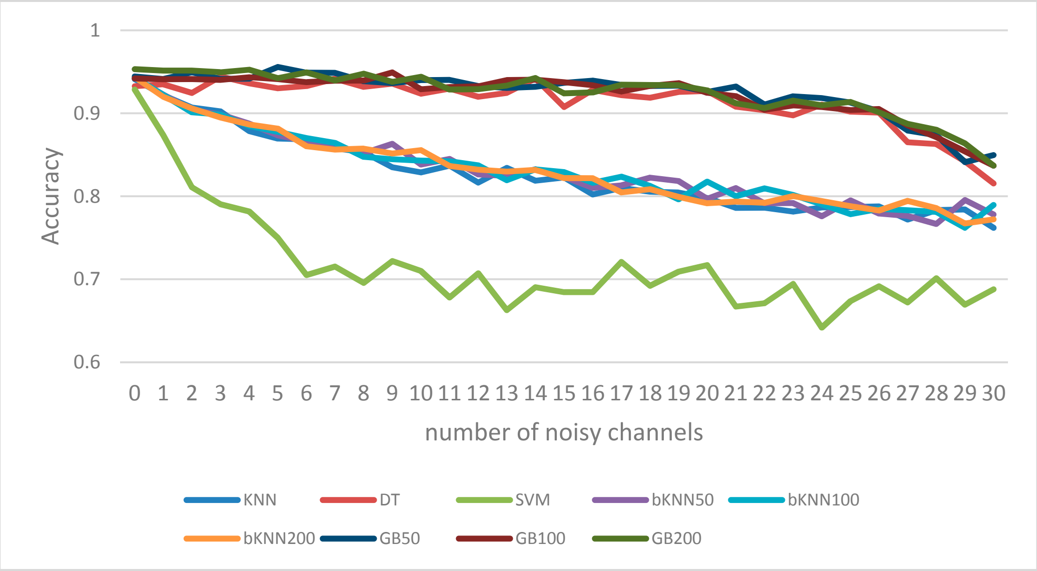


Figure 7. Comparison of different classifiers for impact of noise on detection performance with combined feature sets. The left vertical coordinate is for average accuracy, while the horizontal coordinate is for number of noisy channels.

The above results are summarized in Table 2, and in conjunction with Table 1, it can be seen that combined entropy can enhance robustness.

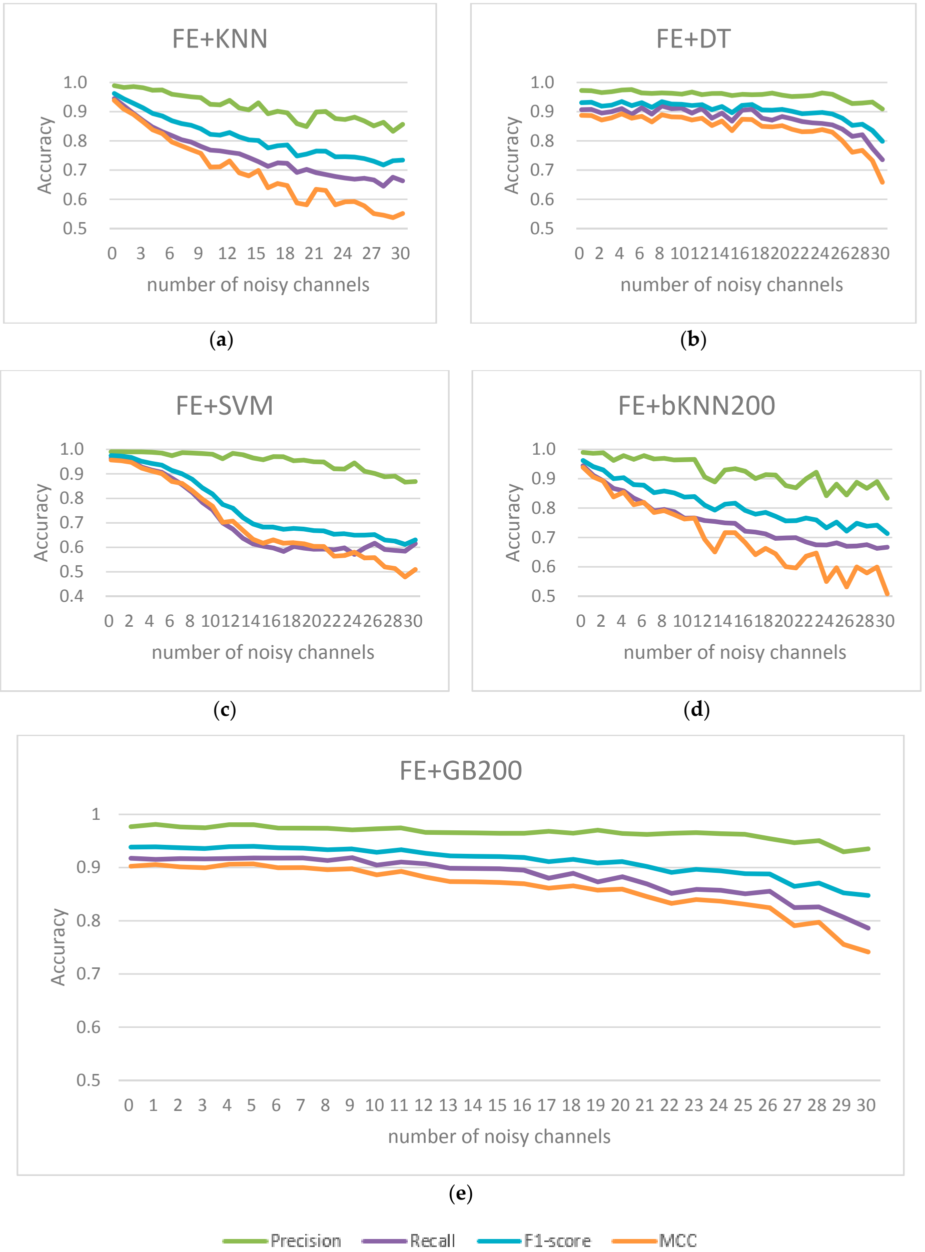
Table 2. Results of the analysis of average accuracy with simulated noise EEG signals for the combined feature set. A_0 is defined as the average accuracy with noise-free signals while A_{30} is defined as the average accuracy with 30 noisy channels signals.

-	A_{30}/A_0
KNN	0.810
DT	0.874
SVM	0.741
bKNN50	0.825
bKNN100	0.837
bKNN200	0.819
GB50	0.900
GB100	0.888
GB200	0.878

3.1.5. Other Performance Indexes

Figure 8 shows a comparison of the different classifiers for the FE feature sets. In this section, Precision, Recall, F1-score and MCC were used as the model performance indicators. A comparison of the results of different prediction methods and FE feature sets indicated that the GB model and DT classifier were statistically different to any of the other techniques, and achieved a better model performance. This finding further confirmed the advantages of the GB model and DT classifier in modeling complex relationships between EEG signals and the fatigue state.

The above results are summarized in Table 3. A_0 is defined as the average accuracy with noise-free signals while A_{30} is defined as the average accuracy with 30 noisy channel signals. In this paper, A_{30}/A_0 was used as an important indicator of robustness. Table 3 summarizes the other four performance indexes of three classifiers and two ensemble methods in the FE feature sets obtained from noisy EEG data. It was noted that the Boosting method had significantly different average accuracies from other methods across all four indexes when EEG data were polluted (paired t-test, $p < 0.01$).



(**a**) (**b**)

(**c**) (**d**)

(**e**)

Figure 8. Comparison of different classifiers on the impact of noise on the detection performance of fuzzy entropy (FE) feature sets. The left vertical coordinate is for average precision, Recall, F1-score and Matthews Correlation Coefficient (MCC), while the horizontal coordinate is for the number of noisy channels. (**a**–**e**) represents classifier KNN, DT, SVM, bKNN200 and GB200, respectively.

Table 3. Results of the analysis of average precision, Recall, F1-score and MCC with simulated noise EEG signals in a 30-channel system. A_0 is defined as the average accuracy with noise-free signals while A_{30} is defined as the average precision, Recall, F1-score and MCC with 30 noisy channel signals.

A_{30}/A_0	Precision	Recall	F1	MCC
KNN	0.866	0.702	0.763	0.587
DT	0.935	0.812	0.858	0.742
SVM	0.878	0.639	0.647	0.533
bKNN200	0.842	0.706	0.742	0.541
GB200	0.957	0.857	0.903	0.822

3.1.6. Effect of Level of Noise

In this section, we used polluted EEG signals that were generated by adding white Gaussian noise with a different scale factor D into the original EEG signal as mentioned in Section 2.6.1. This was accomplished by computing the average accuracy under increasing levels of noise (0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9 and 1.0). The changes to A_{30}/A_0 with levels of noise were investigated in the 30-channel system.

As shown in Figure 9, we found that the classification accuracy of DT was higher than those of the SVM and KNN for all noise levels (paired *t*-test, $p < 0.01$). The difference in classification accuracy between the DT and SVM (KNN) increased with the increase in scale level. Similarly, Figure 10 shows the noise robustness results of the Bagging and Boosting method. It was found that the classification accuracy of DT was higher than that of the SVM (KNN) for all noise levels. In addition, when the noise power increased, the accuracy difference between the DT and SVM increased. For example, in the noiseless case, the average accuracy difference between the SVM and DT was 1.9%; however, in the case of D = 0.5 and 1.0, the difference was 5.8% and 8.5%. These results indicate that the DT method was more robust than the SVM for the polluted EEG signal in the Gaussian white noise case (paired *t*-test, $p < 0.01$). Furthermore, there were no significant differences among the four feature sets (paired *t*-test, $p > 0.05$).

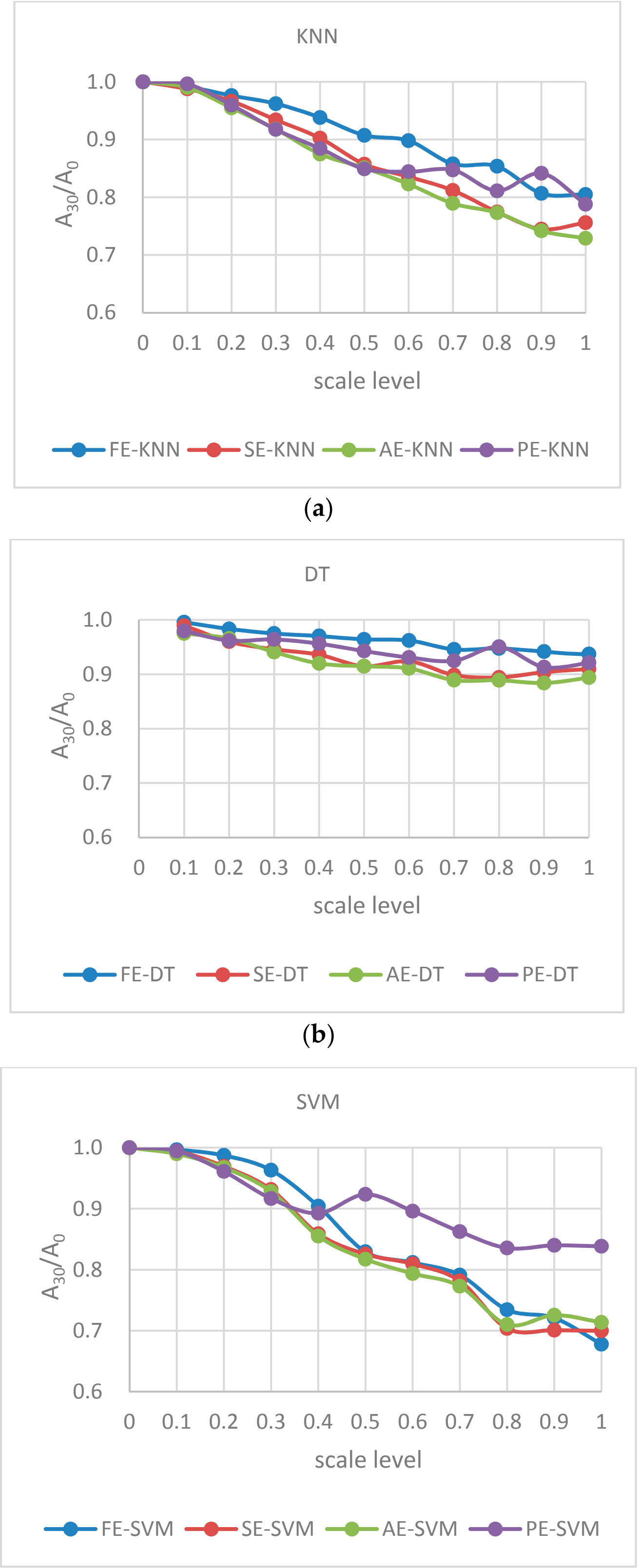
The above results show that the level of noise did not change the effect of noise on the detection performance. Additionally, these results indicate that the Boosting method significantly enhanced the capabilities and robustness of the system, while the Bagging method was unable to do so.

3.2. Spike Noise

The experiment was repeated using spike noise. With a probability of 0.01, and a duration of 1 sample, spikes did not seem to significantly impact the matches count and, therefore, impact on the entropy metrics.

Based on the results of Section 3.1, among the four feature sets, FE performed best, and PE was the worst. Among the three classifiers, DT was the best, and SVM was the weakest. Here, we compare the detection performance for three classifiers and four feature sets with the addition of spike noise.

Figure 11 shows the variation of the average accuracy as a function of the number of noisy channels. Unlike the results described in Section 3.1, with an increase in the number of noisy channels for the five feature sets (FE, SE, AE, PE and Combined), the average accuracy was almost unchanged for different classifiers (paired *t*-test, $p > 0.05$). Given the low frequency of spike noise and the entropy feature extraction method, spike noise had little effect on classification performance. For the four kinds of entropy and the nine classification models, the average accuracy basically changed over a small range.

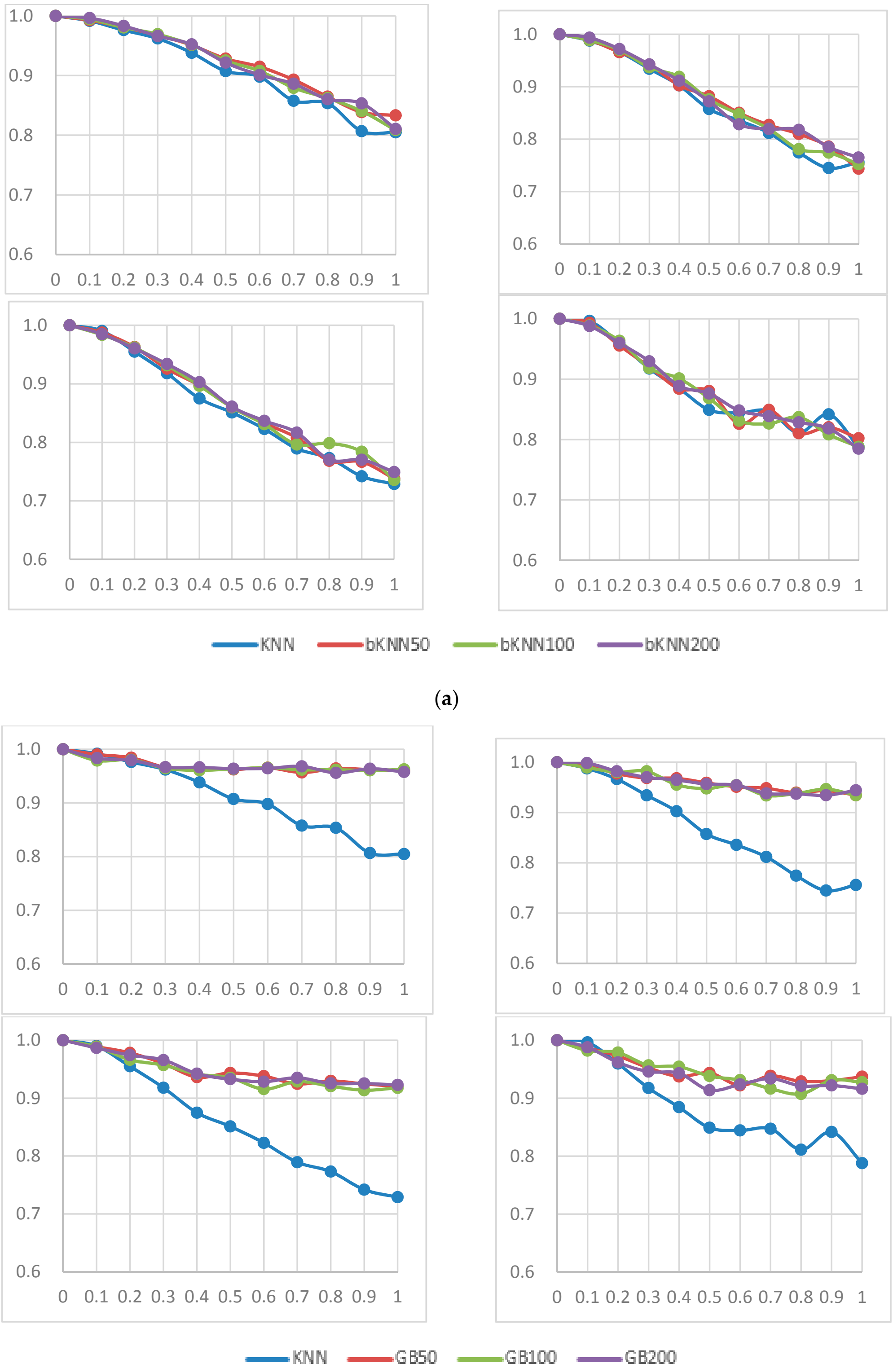


(**a**)

(**b**)

(**c**)

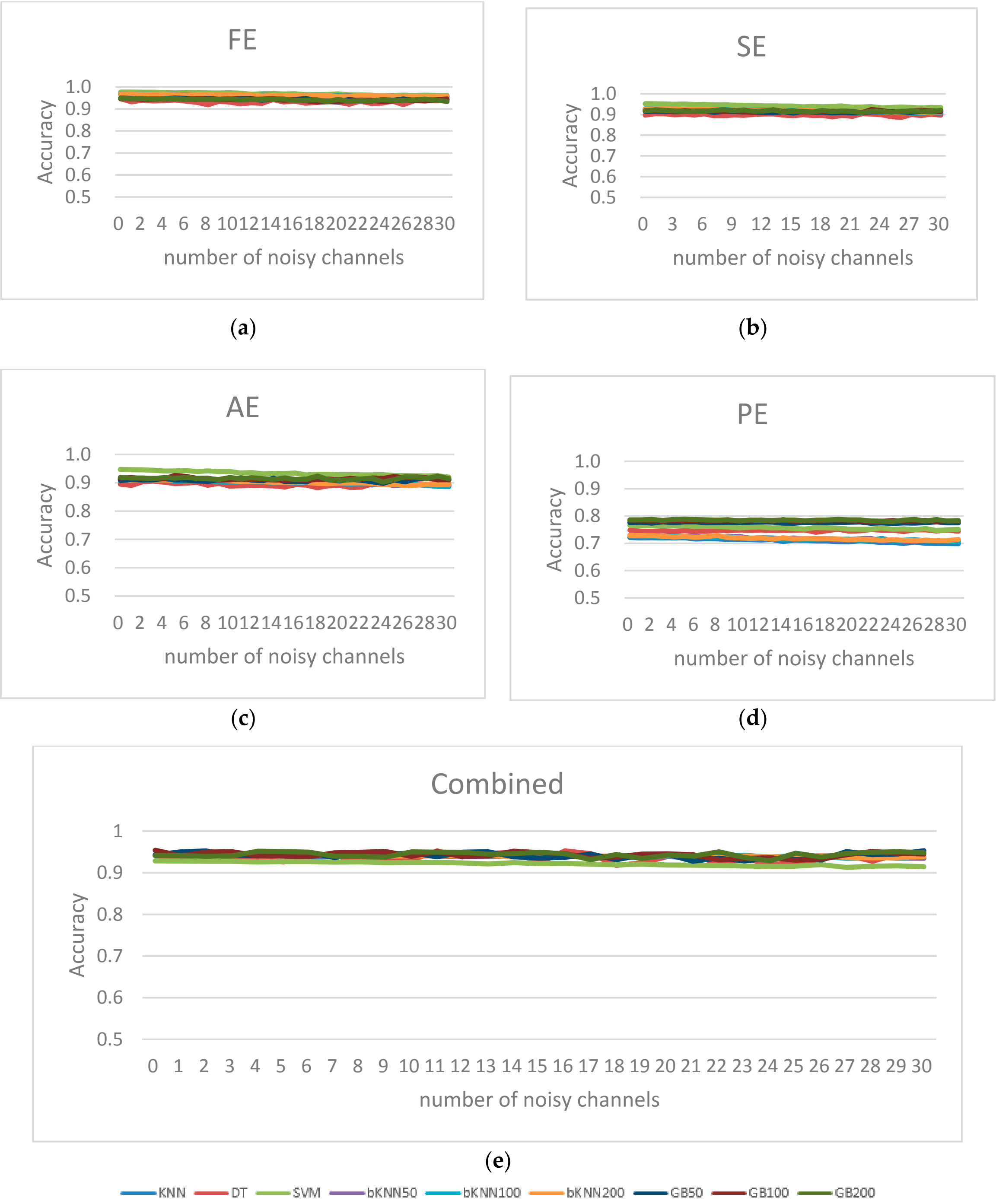
Figure 9. Relationship between A_{30}/A_0 and levels of noise for the three classifiers for (**a**) KNN; (**b**) DT; (**c**) SVM. The left vertical coordinate is the value of A_{30}/A_0, while the horizontal coordinate is the scale level of noise.



(**a**)

(**b**)

Figure 10. Relationship between A_{30}/A_0 and levels of noise based on three classifiers for (**a**) Bagging method and (**b**) Boosting method. The left vertical coordinate is A_{30}/A_0, while the horizontal coordinate is the level of noise. In each subfigure, from top to bottom, from left to right, the results are based on FE, SE, AE, and PE, respectively.



(**a**)

(**b**)

(**c**)

(**d**)

(**e**)

Figure 11. Influence of superimposed noise on the average accuracy for the three classifiers and two ensemble methods based on (**a**) FE feature set; (**b**) SE feature set; (**c**) AE feature set; (**d**) PE feature set and (**e**) combined feature set. The left vertical coordinate is for average accuracy, while the horizontal coordinate is for the number of noisy channels. bKNN50, bKNN100 and bKNN200 represent the Bagging ensemble with 50, 100, and 200 KNN base classifiers. GB50, GB100 and GB200 represent the Boosting ensemble with 50, 100, and 200 Boosting stages.

A_0 is defined as the average accuracy with noise-free signals while A_{30} is defined as the average accuracy with 30 noisy channels signals. In this paper, A_{30}/A_0 was used as an important indicator for robustness. Table 4 summarizes the average accuracy of the three classifiers in four feature sets obtained from noisy EEG data with the addition of spike noise. It was noted that spike noise made no difference to average accuracy across all classifiers and feature sets (paired *t*-test, $p > 0.05$).

Table 4. Analysis of average accuracy with simulated spike noise EEG signals. A_0 is defined as the average accuracy with noise-free signals while A_{30} is defined as the average accuracy with 30 noisy channel signals.

A_{30}/A_0	FE	SE	AE	PE	Combine
KNN	0.988	0.979	0.978	0.968	0.994
DT	0.993	1.000	1.001	0.995	0.992
SVM	0.984	0.980	0.972	0.980	0.985
bKNN50	0.991	0.977	0.978	0.975	0.993
bKNN100	0.988	0.981	0.973	0.972	0.996
bKNN200	0.991	0.985	0.980	0.981	0.996
GB50	0.992	0.999	1.002	0.998	1.010
GB100	1.000	0.998	0.994	0.996	0.991
GB200	0.986	1.003	0.995	0.997	1.007

3.3. EMG Noise

The experiment was repeated using simulated EMG noise.

3.3.1. Effect of Noise: DT Classifier vs. KNN Classifier vs. SVM Classifier

Based on the results described in Section 3.1, among the four feature sets, FE performed best, and PE was the worst. Among the three classifiers, DT was the best, and SVM was the weakest. Here we compare the detection performance for the three classifiers and four feature sets with EMG noise.

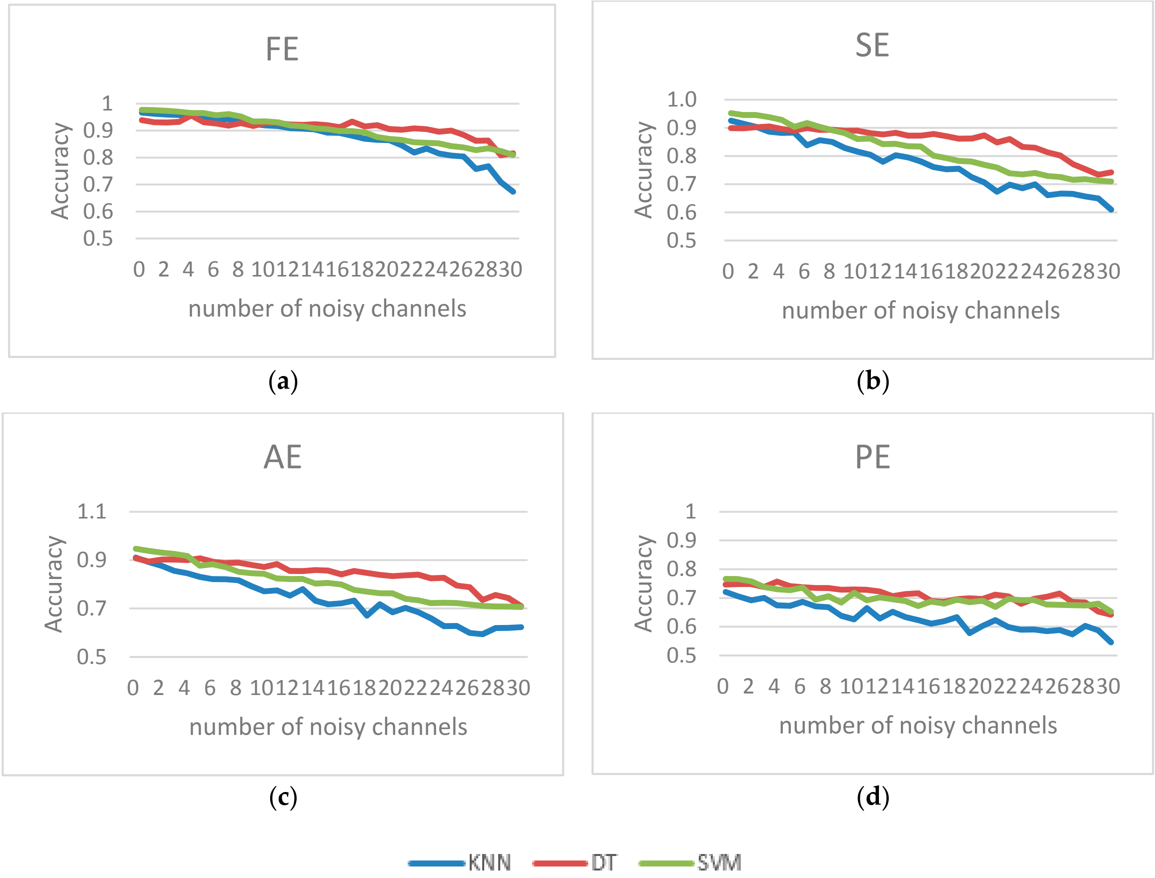
Figure 12 shows the variation of the average accuracy as a function of the number of noisy channels. For the four feature sets (FE, SE, AE, and PE), the difference between the various classifiers for the noise-free EEG signal and noisy EEG signal was clear and remained consistent, with a greater performance of the DT classifier than those of the SVM and KNN classifiers. However, unlike the results seen in Section 3.1, DT decreased significantly.

When using the feature set FE, there were no differences in average accuracy between the KNN, DT, and SVM classifiers for the original EEG signals (paired *t*-test, $p > 0.05$). However, with more channels superimposing noise, the average accuracy of the three classifiers decreased. For the DT classifier, the average accuracy decreased from 0.939 with noise-free signals to 0.816 with 30 noisy channels. For the SVM and KNN, the average accuracies were 0.976 and 0.966 with noise-free signal and dropped to 0.810 and 0.673 with 30 noisy channels, respectively. The effect of DT and SVM was better than that of KNN in the presence of 30 noisy channels (paired *t*-test, $p < 0.01$).

When using the feature set AE, there were no differences in average accuracy among the KNN, DT, and SVM for the original EEG signals (paired *t*-test, $p > 0.05$). However, with more channels adding noise, the average accuracy of three classifiers decreased. For the DT classifier, the average accuracy decreased from 0.899 with noise-free signal to 0.742 with 30 noisy channels. For the SVM and KNN, the average accuracies were 0.952 and 0.925 with noise-free signals and dropped to 0.709 and 0.610 with 30 noisy channels, respectively. Therefore, the effects of DT were better than those of the SVM and KNN classifiers in the presence of 30 noisy channels (paired *t*-test, $p < 0.01$), and SE was similar to AE.

When using the feature set PE, there were no differences in the average accuracy among the KNN, DT and SVM for the original EEG signals (paired *t*-test, $p > 0.05$). However, with more channels adding noise, the average accuracy of the three classifiers decreased until the final average accuracy was almost the same. For the SVM classifier, the average accuracy decreased from 0.766 with noise-free signal to 0.652 with 30 noisy channels. For the DT and KNN, the average accuracies were 0.746 and 0.721 with noise-free signal and dropped to 0.642 and 0.546 with 30 noisy channels, respectively.

The above results show that: (1) when EMG signals were superimposed; the effect of noise was greater than that of the Gaussian noise; (2) Similarly, while the FE feature set had higher robustness, AE and SE were similar. The average accuracy for the PE feature set with original noise-free EEG was not high, and was lower than those of the FE, SE, and AE feature sets significantly (paired *t*-test, $p < 0.01$), so there was little difference between the three classifiers. Finally, DT had the best robustness, and SVM and KNN were similar.



(**a**) (**b**)

(**c**) (**d**)

Figure 12. Influence of superimposed noise on the average accuracy for the three classifiers when using (**a**) FE feature set; (**b**) SE feature set; (**c**) AE feature set and (**d**) PE feature set. The left vertical coordinate is for average accuracy, while the horizontal coordinate is for the number of noisy channels.

3.3.2. Effect of Noise: Using Bagging Ensemble Learning Method

Like Section 3.1.2, the average accuracy for the Bagging method decreased the same as that for KNN without the Bagging method when the noisy channels increased (paired *t*-test, $p > 0.05$). As shown in Figure 13, when using the feature set FE, there were no differences in average accuracy between the KNN without the Bagging method and KNN with the Bagging method (paired *t*-test, $p > 0.05$). For the KNN classifier, the average accuracy decreased from 0.966 with a noise-free signal to 0.673 with 30 noisy channels. For the Bagging method with 50, 100, and 200 base classifiers, the average accuracies were 0.965, 0.966 and 0.966 with the noise-free signals and dropped to 0.688, 0.693 and 0.692 with 30 noisy channels, respectively. The other three feature sets were similar to FE.

The above results show that the Bagging method could not significantly improve the recognition effect of the KNN classifier without noisy channels and with noisy channels, and there were no obvious effects when the number of base classifiers was increased.

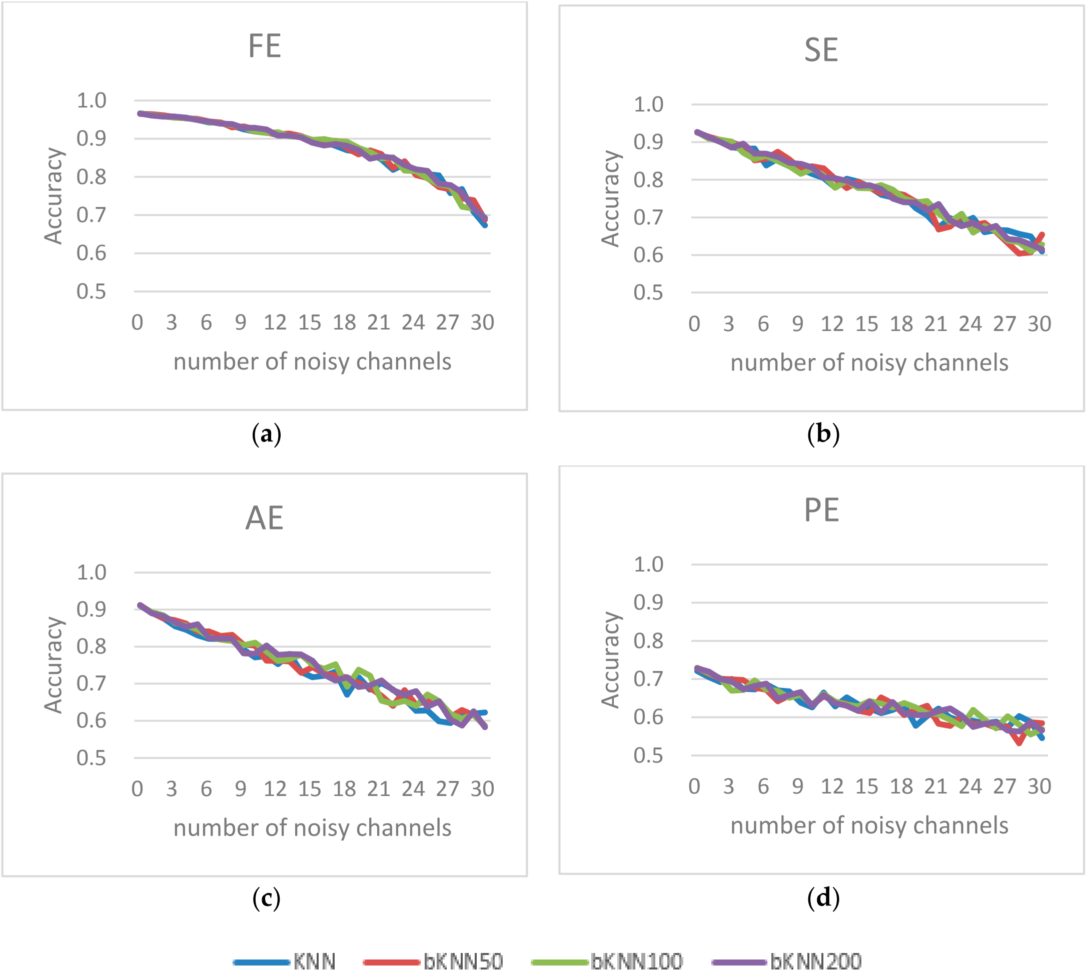


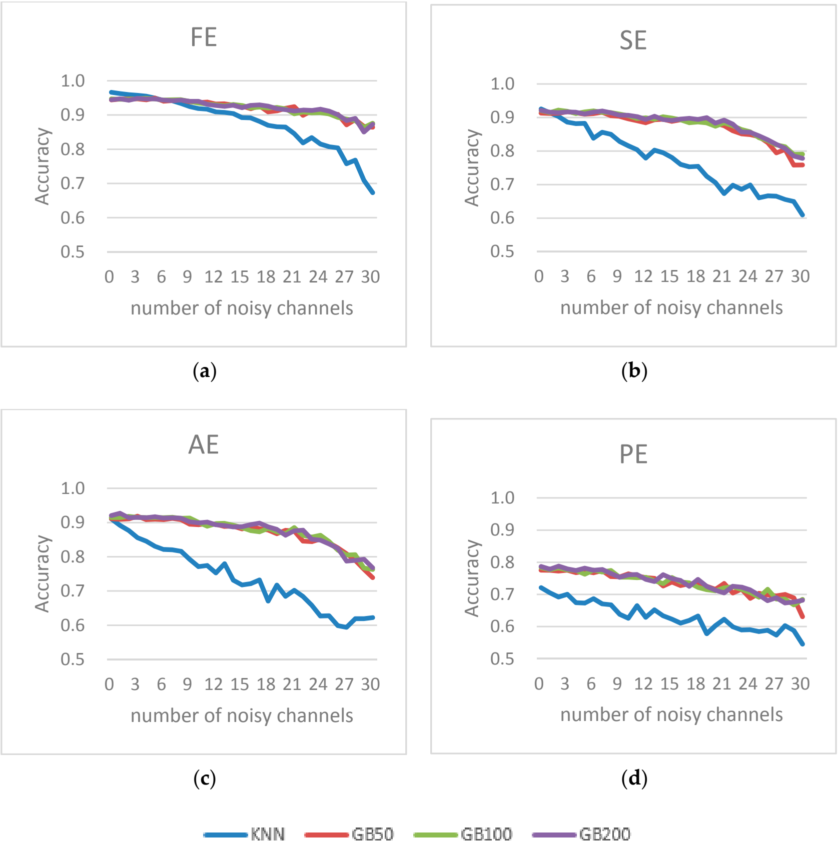
Figure 13. Influence of added noise on the average accuracy for three classifiers with the Bagging method when using (**a**) FE feature set; (**b**) SE feature set; (**c**) AE feature set and (**d**) PE feature set. The left vertical coordinate is for average accuracy, while the horizontal coordinate is for the number of noisy channels. bKNN50, bKNN100 and bKNN200 represent the Bagging ensemble with 50, 100, and 200 KNN base classifiers.

3.3.3. Effect of Noise: Using Boosting Ensemble Learning Method

As shown in Figure 9, in the case of the Boosting method, for four feature sets (FE, SE, AE, and PE), the difference between various classifiers for the noise-free EEG signal and noisy EEG signal was clear and remained consistent. The average accuracy for the Boosting method decreased slower than that for KNN without the Boosting method when the noisy channels increase.

As shown in Figure 14, when using the feature set FE, there was a difference in average accuracy between the KNN and Boosting method (paired t-test, $p < 0.01$). With more channels adding noise, the average accuracy of both classifiers decreased, but the Boosting method decreased slower. For the KNN classifier, the average accuracy decreased from 0.966 with a noise-free signal to 0.673 with 30 noisy channels (paired t-test, $p < 0.01$). For the Boosting method with 50, 100, and 200 base classifiers, the average accuracies were 0.944, 0.947 and 0.945 with noise-free signals and dropped to 0.864, 0.877 and 0.873 with 30 noisy channels, respectively. The other three feature sets were similar to FE.

The above results show that the Boosting method could not effectively improve the recognition effect of the KNN classifier without noisy channels; however, it could significantly improve the recognition effect of KNN classifiers with noisy channels. Additionally, there was no obvious effect when the number of base classifiers was increased.



(**a**) (**b**) (**c**) (**d**)

Figure 14. Influence of added noise on the average accuracy for the three classifiers with the Boosting method when using (**a**) FE feature set; (**b**) SE feature set; (**c**) AE feature set and (**d**) PE feature set. The left vertical coordinate is for average accuracy, while the horizontal coordinate is for the number of noisy channels. GB50, GB100 and GB200 represent the Boosting ensemble with 50, 100, and 200 Boosting stages.

A_0 is defined as the average accuracy with noise-free signals while A_{30} is defined as the average accuracy with 30 noisy channels signals. In this paper, A_{30}/A_0 was used as an important indicator for robustness. Table 5 summarizes the average accuracy of the three classifiers in four feature sets obtained from noisy EEG data in a 30-channel system. It was noted that the Boosting method had significantly different average accuracies from the KNN method across all feature sets when the EEG data were preprocessed. In addition, FE achieved a better performance than those of AE and SE.

Table 5. Analysis of average accuracy with simulated noise EEG signals. A_0 is defined as the average accuracy with noise-free signals while A_{30} is defined as the average accuracy with 30 noisy channel signals.

A_{30}/A_0	FE	SE	AE	PE	Combined
KNN	0.697	0.659	0.683	0.756	0.890
DT	0.869	0.825	0.781	0.860	0.957
SVM	0.830	0.745	0.745	0.852	0.825
bKNN50	0.713	0.707	0.643	0.803	0.884
bKNN100	0.718	0.678	0.644	0.780	0.901
bKNN200	0.717	0.664	0.639	0.777	0.892
GB50	0.916	0.831	0.811	0.814	0.981
GB100	0.925	0.859	0.834	0.873	0.976
GB200	0.924	0.845	0.835	0.866	0.979

As shown in Figure 15, there were no differences in average accuracy among KNN, DT, and SVM for the unpolluted EEG signals (paired t-test, $p > 0.05$). However, with more channels adding EMG noise, the average accuracy of the three classifiers decreased, but DT decreased slower. For the DT classifier, the average accuracy decreased from 0.932 with a noise-free signal to 0.892 with 30 noisy channels. For the SVM and KNN, the average accuracies were 0.929 and 0.941 with noise-free signals and dropped to 0.766 and 0.838 with 30 noisy channels, respectively. The effect of DT was better than those of the SVM and KNN classifiers in the presence of 30 noisy channels (paired t-test, $p < 0.01$). For the Bagging method with 50, 100, and 200 base classifiers, the average accuracies were 0.943, 0.943, and 0.943 with noise-free signals and dropped to 0.833, 0.849, and 0.841 with 30 noisy channels, respectively (paired t-test, $p > 0.05$). The above results showed that the Bagging method could not effectively improve the recognition effect of the KNN classifier without noisy channels and with noisy channels. There was also no obvious effect when the number of base classifiers was increased. For the Boosting method with 50, 100, and 200 base classifiers, the average accuracies were 0.941, 0.952, and 0.953 with noise-free signals and dropped to 0.923, 0.929, and 0.933 with 30 noisy channels, respectively (paired t-test, $p < 0.01$). The above results showed that the Boosting method was unable to improve the recognition effect without noisy channels; however, it could significantly improve the recognition effect with noisy channels. There was no obvious effect when increasing the number of base classifiers.

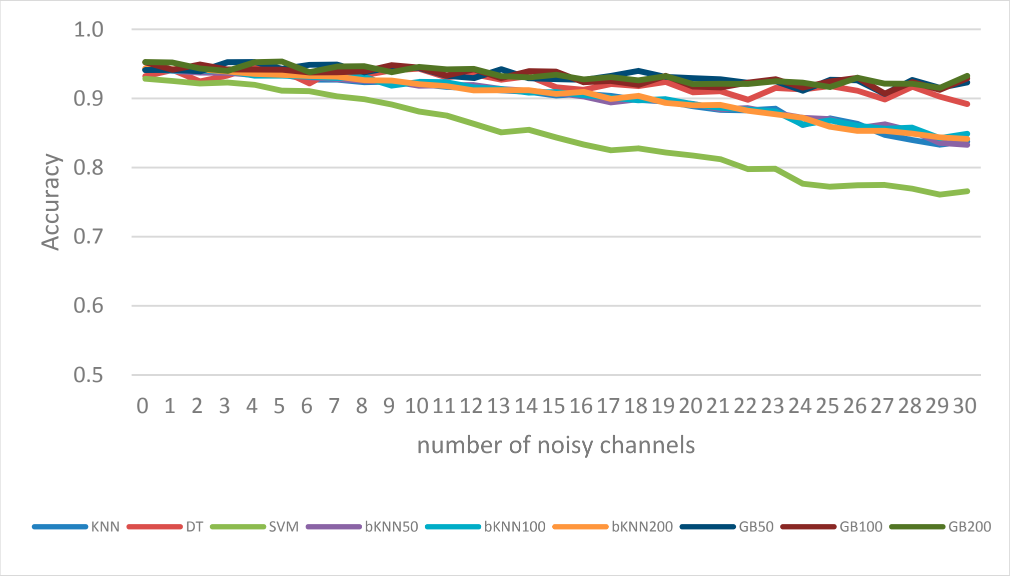


Figure 15. Comparison of different classifiers for the impact of noise on detection performance with combined feature sets. The left vertical coordinate is for average accuracy, while the horizontal coordinate is for the number of noisy channels.

4. Conclusions

In this study, an approach based on simulated Gaussian noise was proposed to investigate the effect of different classifiers and four feature sets in detecting driver fatigue in an EEG-based system. For this purpose, we generated noise corrupted EEG signals using simulated Gaussian noise, Spike noise and simulated EMG noise. Next, we assessed the detection performance of various classifier methods with a varied number of noisy channels. Using the experimental driver fatigue based EEG and generated noisy signals, we compared the classification results of the DT, SVM, and KNN methods. From our results, it was evident that DT showed superior noise robustness than the SVM and KNN methods. Furthermore, the results showed that the classification accuracy of FE and the combined feature set were better than those of the other feature sets. It was also found that the Bagging method could not effectively improve performance with noise, while the Boosting method may have effectively improved performance with noise.

Practically, the proposed method may face more problems outside the EEG acquisition from the lab. One of the most important is the noise issue as there are many artifacts that may affect driving fatigue recognition. Currently, there has been some research focused on artifact removal methods prior to the feature extraction process, but these methods may also cause problems in the elimination of the artifacts, and also weaken the feature, such as the average method. In addition, it may lead to computational complexity and temporal extension, which is unfavorable in practical applications. This study revealed that the extraction method with an appropriate combination of entropy features (such as FE or combined feature sets) and classifier (such as DT or Boosting) could not only improve the recognition rate; but could weaken the noise impact on the recognition rate.

However, some limitations of this study are: (1) the number of subjects was relatively small. Although the existing literature suggests that 22 subjects is not too small a sample size, the number still needs to be increased; (2) Only three commonly used classifiers and the four feature sets were compared in this study; (3) For simplicity, the noise and the original signal were subject to linear superposition. However, the models of external noise were diverse, and the interaction model with the original EEG signal were also diverse. Finally, the different impacts of different channels were not considered.

Acknowledgments: This work was supported by Project of Department of Science and Technology, Jiangxi Province (No. 20151BBE50079), and Project of Department of Education, Jiangxi Province (No. GJJ151146 and No. GJJ161143). Thanks Jian-Liang Min and Zhen-Dong Mu for collecting and preprocessing EEG data.

Author Contributions: Jianfeng Hu conceived and designed the experiments; Ping Wang performed the experiments and analyzed the data; all authors wrote the paper. All the authors have read and approved the final manuscript.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Correa, A.G.; Orosco, L.; Laciar, E. Automatic detection of drowsiness in EEG records based on multimodal analysis. *Med. Eng. Phys.* **2014**, *36*, 244–249. [CrossRef] [PubMed]
2. Mu, Z.D.; Hu, J.F.; Yin, J.H. Driving Fatigue Detecting Based on EEG Signals of Forehead Area. *Int. J. Pattern Recognit. Artif. Intell.* **2017**, *31*, 1750011. [CrossRef]
3. Mu, Z.D.; Hu, J.F.; Min, J.L. Driver Fatigue Detection System Using Electroencephalography Signals Based on Combined Entropy Features. *Appl. Sci.* **2017**, *7*, 150. [CrossRef]
4. Fu, R.R.; Wang, H.; Zhao, W.B. Dynamic driver fatigue detection using hidden Markov model in real driving condition. *Expert Syst. Appl.* **2016**, *63*, 397–411. [CrossRef]
5. Li, W.; He, Q.C.; Fan, X.M.; Fei, Z.M. Evaluation of driver fatigue on two channels of EEG data. *Neurosci. Lett.* **2012**, *506*, 235–239. [CrossRef] [PubMed]
6. Xiong, Y.; Gao, J.; Yang, Y.; Yu, X.; Huang, W. Classifying Driving Fatigue Based on Combined Entropy Measure Using EEG Signals. *Int. J. Control Autom.* **2016**, *9*, 329–338. [CrossRef]
7. Chai, R.; Naik, G.; Nguyen, T.N.; Ling, S.; Tran, Y.; Craig, A.; Nguyen, H. Driver fatigue classification with independent component by entropy rate bound minimization analysis in an EEG-based system. *IEEE J. Biomed. Health Inform.* **2016**. [CrossRef] [PubMed]
8. Chai, R.; Ling, S.H.; San, P.P.; Naik, G.R.; Nguyen, T.N.; Tran, Y.; Craig, A.; Nguyen, H.T. Improving EEG-Based Driver Fatigue Classification Using Sparse-Deep Belief Networks. *Front. Neurosci.* **2017**, *11*, 103. [CrossRef] [PubMed]
9. Chai, R.; Naik, G.R.; Tran, Y.; Ling, S.H.; Craig, A.; Nguyen, H.T. Classification of driver fatigue in an electroencephalography-based countermeasure system with source separation module. In Proceedings of the 37th Annual International Conference of the IEEE on Engineering in Medicine and Biology Society, Milan, Italy, 25–29 August 2015; pp. 514–517.

10. Wu, D.; Lawhern, V.J.; Gordon, S.; Lance, B.J.; Lin, C.T. Driver Drowsiness Estimation from EEG Signals Using Online Weighted Adaptation Regularization for Regression. *IEEE Trans. Fuzzy Syst.* **2017**, *99*, 1. [CrossRef]
11. Huang, K.C.; Huang, T.Y.; Chuang, C.H.; King, J.T.; Wang, Y.K.; Lin, C.T.; Jung, T.P. An EEG-based fatigue detection and mitigation system. *Int. J. Neural Syst.* **2016**, *26*, 1650018. [CrossRef] [PubMed]
12. Hassan, A.R.; Bhuiyan, M.I.H. Computer-Aided Sleep Staging Using Complete Ensemble Empirical Mode Decomposition with Adaptive Noise and Bootstrap Aggregating. *Biomed. Signal Process. Control* **2016**, *24*, 1–10. [CrossRef]
13. Hassan, A.R.; Subasi, A. Automatic Identification of Epileptic Seizures from EEG Signals Using Linear Programming Boosting. *Comput. Methods Progr. Biomed.* **2016**, *136*, 65–77. [CrossRef] [PubMed]
14. Sun, S.L.; Zhang, C.S.; Zhang, D. An experimental evaluation of ensemble methods for EEG signal classification. *Pattern Recognit. Lett.* **2007**, *28*, 2157–2163. [CrossRef]
15. Yang, T.; Chen, W.T.; Cao, G.T. Automated Classification of Neonatal Amplitude-Integrated EEG Based on Gradient Boosting Method. *Biomed. Signal Process. Control* **2016**, *28*, 50–57. [CrossRef]
16. Sanei, S.; Chambers, J.A. *EEG Signal Processing*; Wiley: New York, NY, USA, 2007.
17. Lotte, F.; Congedo, M.; Lecuyer, A.; Lamarche, F.; Arnaldi, B. A review of classification algorithms for EEG-based brain–computer interfaces. *J. Neural Eng.* **2007**, *4*, 1–13. [CrossRef] [PubMed]
18. Islam, M.K.; Rastegarni, A.; Yang, Z. Methods for artifact detection and removal from scalp EEG: A review. *Clin. Neurophysiol.* **2016**, *46*, 287–305. [CrossRef] [PubMed]
19. Minguillon, J.; Lopez-Gordo, M.A.; Pelayo, F. Trends in EEG-BCI for daily-life: Requirements for artifact removal. *Biomed. Signal Process. Control* **2017**, *31*, 407–418. [CrossRef]
20. Azarnoosh, M.; Nasrabadi, A.M.; Mohammadi, M.R.; Firoozabadi, M. Investigation of mental fatigue through EEG signal processing based on nonlinear analysis: Symbolic dynamics. *Chaos Solitons Fractals* **2011**, *44*, 1054–1062. [CrossRef]
21. Kannathal, N.; Choo, M.L.; Acharya, U.R.; Sadasivan, P. Entropies for detection of epilepsy in EEG. *Comput. Methods Progr. Biomed.* **2005**, *80*, 187–194. [CrossRef] [PubMed]
22. Xiang, J.; Li, C.; Li, H.; Cao, R.; Wang, B.; Han, X.; Chen, J. The detection of epileptic seizure signals based on fuzzy entropy. *J. Neurosci. Methods* **2015**, *243*, 18–25. [CrossRef] [PubMed]
23. Pincus, S.M. Approximate entropy as a measure of system complexity. *Proc. Natl. Acad. Sci. USA* **1991**, *88*, 2297–2301. [CrossRef] [PubMed]
24. Richman, J.S.; Moorman, J.R. Physiological time-series analysis using approximate entropy and sample entropy. *Am. J. Physiol. Heart Circ. Physiol.* **2000**, *278*, 2039–2049.
25. Chen, W.; Zhuang, J.; Yu, W.; Wang, Z. Measuring complexity using fuzzyen, apen, and sampan. *Med. Eng. Phys.* **2009**, *31*, 61–68. [CrossRef] [PubMed]
26. Lee, K.A.; Hicks, G.; Nino-Murcia, G. Validity and reliability of a scale to assess fatigue. *Psychiatry Res.* **1991**, *36*, 291–298. [CrossRef]
27. Borg, G. Psychophysical scaling with applications in physical work and the perception of exertion. *Scand. J. Work Environ. Health* **1990**, *16*, 55–58. [CrossRef] [PubMed]
28. Song, Y.; Crowcroft, J.; Zhang, J. Automatic epileptic seizure detection in EEGs based on optimized sample entropy and extreme learning machine. *J. Neurosci. Methods* **2012**, *210*, 132–146. [CrossRef] [PubMed]
29. Yentes, J.M.; Hunt, N.; Schmid, K.K.; Kaipust, J.P.; McGrath, D.; Stergiou, N. The appropriate use of approximate entropy and sample entropy with short data sets. *Ann. Biomed. Eng.* **2013**, *41*, 349–365. [CrossRef] [PubMed]
30. Breiman, L. Bagging predictors. *Mach. Learn.* **1996**, *24*, 123–140. [CrossRef]
31. Breiman, L. Heuristics of instability and stabilization in model selection. *Ann. Stat.* **1996**, *24*, 2350–2383. [CrossRef]
32. Friedman, J.H. Greedy Function Approximation: A Gradient Boosting Machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [CrossRef]
33. Hastie, T.; Tibshirani, R.; Friedman, J.H. *Elements of Statistical Learning*; Springer: Berlin, Germany, 2009.
34. Dietterich, T.G. An experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization. *Mach. Learn.* **2000**, *40*, 139–157. [CrossRef]
35. Bauer, E.; Kohavi, R. An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants. *Mach. Learn.* **1999**, *36*, 105–139. [CrossRef]

36. Cawley, G.C.; Talbot, N.L.C. On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. *JMLR* **2010**, *11*, 2079–2107.
37. Hu, J.F. Comparison of Different Features and Classifiers for Driver Fatigue Detection Based on a Single EEG Channel. *Comput. Math. Methods Med.* **2017**. [CrossRef] [PubMed]
38. Mu, Z.D.; Hu, J.F.; Min, J.L. EEG-Based Person Authentication Using a Fuzzy Entropy-Related Approach with Two Electrodes. *Entropy* **2016**, *18*, 432. [CrossRef]
39. Yin, J.H.; Hu, J.F.; Mu, Z.D. Developing and evaluating a Mobile Driver Fatigue Detection Network Based on Electroencephalograph Signals. *Healthc. Technol. Lett.* **2017**, *4*, 34–38. [CrossRef] [PubMed]
40. Hu, J.F.; Mu, Z.D.; Wang, P. Multi-feature authentication system based on event evoked electroencephalogram. *J. Med. Imaging Health Inform.* **2015**, *5*, 862–870.
41. Hu, J.F. Automated detection of driver fatigue based on AdaBoost classifier with EEG signals. *Front. Comput. Neurosci.* **2017**. [CrossRef]
42. Mu, Z.D.; Hu, J.F.; Min, J.L.; Yin, J.H. Comparison of Different Entropy as Feature for Person Authentication Based on EEG Signals. *IET Biom.* **2017**. [CrossRef]
43. Srivastava, N.; Hinton, G.E.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929–1958.
44. Quinlan, J.R. Bagging, Boosting, and C4.5. Available online: https://pdfs.semanticscholar.org/79ea/6a5a68e05065f82acd11a478aa7eac5f6c06.pdf (accessed on 25 July 2017).
45. Cuesta-Frau, D.; Aboy, M.; Crespo, C.; Oltra-Crespo, S. Comparative study of approximate entropy and sample entropy robustness to spikes. *Artif. Intell. Med.* **2011**, *53*, 97–106.
46. Goldberger, A.L.; Amaral, L.A.N.; Glass, L.; Hausdor, J.M.; Ivanov, P.C.; Mark, R.G.; Mietus, J.E.; Moody, G.B.; Peng, C.K.; Stanley, H.E. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation* **2000**, *101*, 215–220. [CrossRef]
47. Sokolova, M.; Japkowicz, N.; Szpakowicz, S. Beyond accuracy, F-score and ROC: A family of discriminant measures for performance evaluation. In Proceedings of the Australasian Joint Conference on Artificial Intelligence, Hobart, Australia, 4–8 December 2006; pp. 1015–1021.