applied sciences

MDPI

Article

Evaluation of Unsupervised Anomaly Detection Techniques in Labelling Epileptic Seizures on Human EEG

Oleg E. Karpov [1], Matvey S. Khoymov [2], Vladimir A. Maksimenko [2], Vadim V. Grubov [2], Nikita Utyashev [1], Denis A. Andrikov [3], Semen A. Kurkin [2] and Alexander E. Hramov [2,*]

[1] National Medical and Surgical Center named after N.I. Pirogov, Ministry of Healthcare of the Russian Federation, 105203 Moscow, Russia
[2] Baltic Center for Neurotechnology and Artificial Intelligence, Immanuel Kant Baltic Federal University, 236041 Kaliningrad, Russia
[3] Research and Production Company "Immersmed", 105203 Moscow, Russia
* Correspondence: aekhramov@kantiana.ru

Abstract: Automated labelling of epileptic seizures on electroencephalograms is an essential interdisciplinary task of diagnostics. Traditional machine learning approaches operate in a supervised fashion requiring complex pre-processing procedures that are usually labour intensive and time-consuming. The biggest issue with the analysis of electroencephalograms is the artefacts caused by head movements, eye blinks, and other non-physiological reasons. Similarly to epileptic seizures, artefacts produce rare high-amplitude spikes on electroencephalograms, complicating their separability. We suggest that artefacts and seizures are rare events; therefore, separating them from the rest data seriously reduces information for further processing. Based on the occasional nature of these events and their distinctive pattern, we propose using anomaly detection algorithms for their detection. These algorithms are unsupervised and require minimal pre-processing. In this work, we test the possibility of an anomaly (or outlier) detection algorithm to detect seizures. We compared the state-of-the-art outlier detection algorithms and showed how their performance varied depending on input data. Our results evidence that outlier detection methods can detect all seizures reaching 100% recall, while their precision barely exceeds 30%. However, the small number of seizures means that the algorithm outputs a set of few events that could be quickly classified by an expert. Thus, we believe that outlier detection algorithms could be used for the rapid analysis of electroencephalograms to save the time and effort of experts.

Keywords: epilepsy; electroencephalogram; machine learning; extreme events; anomaly detection; unsupervised

check for updates

Citation: Karpov, O.E.; Khoymov, M.S.; Maksimenko, V.A.; Grubov, V.V.; Utyashev, N.; Andrikov, D.A.; Kurkin, S.A.; Hramov, A.E. Evaluation of Unsupervised Anomaly Detection Techniques in Labelling Epileptic Seizures on Human EEG. *Appl. Sci.* **2023**, *13*, 5655. https://doi.org/10.3390/app13095655

Academic Editor: Francesco Cappello

Received: 22 March 2023
Revised: 1 May 2023
Accepted: 2 May 2023
Published: 4 May 2023

1. Introduction

Epilepsy is a neurological disease that affects about 1% of the world's population [1]. A key feature of epilepsy is seizures—rare episodes of abnormal activity [2]. On the behavioural level, seizures may either induce uncontrollable convulsions (e.g., tonic–clonic seizures) or loss of consciousness (e.g., absence seizures) [3]. On the neural activity level, seizures are accompanied by rhythmic low-frequency neural oscillations with a high amplitude reflecting synchronous activity in large neuronal populations [4].

In general, seizures are accompanied by a state of incapacity that leads to dangerous situations for patients and other people. Epilepsy can also cause cognitive and behavioural deficits [5]. Thus, antiepileptic treatment is highly important, and it starts with diagnostics [6]. Epileptic seizures occur unpredictably and alternate with long periods of normal activity. The rare nature of these events complicates their diagnostics in hospital settings. To collect a sufficient amount of epileptiform activity, patients undergo prolonged electroencephalogram (EEG) monitoring in hospital [7]. As a result, this procedure generates big data comprising hours of normal activity and several seizures with a total duration of a few minutes in total. Usually, this EEG data becomes a subject of manual processing by epileptologists, who spend many hours inspecting EEG signals to label the fragments containing seizures [8]. Machine learning (ML) algorithms are often implemented in EEG data processing, which includes seizure detection [9–12]. These algorithms can be used to reduce the routine work of experts, helping them label seizures faster, and decreasing the number of errors caused by the human factor.

The application of the ML approach in seizure labelling commonly results in classifiers capable of detecting two classes: "seizures" and "normal activity" [13]. Each classifier falls into one of two broad categories: supervised and unsupervised [14]. The supervised ML algorithm undergoes training using the labelled data of some patients before labelling data from a new patient. Using EEG recordings with manually labelled seizures, these algorithms may learn the distinctive features of epileptiform patterns to detect them on unseen EEG signals [15]. A supervised approach to epilepsy detection has shown promising results and includes state-of-the-art techniques such as deep learning (DL) [16], recurrent neural networks (RNNs) [17], and its modified version—long short-term memory (LSTM) [18]. The majority of ML methods for seizure labelling are supervised [19]. However, some limitations of supervised ML algorithms are often overlooked. Firstly, supervised methods require a large amount of pre-labelled data. The acquisition of such a dataset can be a challenging task in itself. Epileptic patterns are known to be high variable, originating from physiological features of different types of epilepsy and further exacerbated by differences in experimental conditions, and hard- and software acquisition. Secondly, one of the most prominent problems in epileptic data is class imbalance: hour and day recordings usually only contain several epileptic episodes with a total length in minutes. In many cases, the class imbalance leads to overfitting.

Unsupervised ML algorithms are commonly less effective in classification but can provide other benefits, for instance, when working with imbalanced data [20]. Additionally, unsupervised methods do not require data pre-labelling and tend to be more explainable [21]. The ultimate goal of our work was to propose a decision support system (DSS) for epilepsy diagnostics as explainability is highly important in medicine. Thus, we decided to use unsupervised ML methods specifically designed to work with imbalanced data and perform anomaly detection. A literature search showed that some unsupervised methods rely on complex pre-processing [22] or invasive techniques [23], neither of which are acceptable for rapid clinical diagnostics in human patients. In this research, we decided to use EEG data with minimal pre-processing to devise a system capable of operating in a clinical setting. The main goal of our work was to estimate the performance of unsupervised anomaly detection models as a system for automatic pre-labelling of epileptic seizures on human EEGs.

Recently, we showed that EEG data with epileptic seizures are not only imbalanced but also demonstrate features of extreme event behaviour [24,25]. We used this knowledge to apply anomaly detection techniques to label epileptic EEG data. These techniques belong to unsupervised ML algorithms that operate without training data. We tested a one-class support-vector machine (SVM), a popular outlier detection algorithm, on the data of 83 patients and reported 77% sensitivity and 12% precision [26].

Here, we further investigate the ability of outlier detection algorithms to detect seizures. We start by introducing popular algorithms for outlier detection, including SVM, ensemble method, and distance-based methods, and adjust their hyperparameters to ensure the best score. Then, we trained these algorithms using various features of EEG signals, including two frequency ranges and three types of spectral power: full spectrum, mean, and principal components (PCs). With the obtained results, we tested two hypotheses: (i) the algorithm type and features impact performance; (ii) predicted performance based on prior knowledge about the algorithm type and features.

The remainder of this paper is organized as follows. Section 2 presents the materials and methods, describing the used dataset, details on the following data processing, and feature engineering procedures. Section 2 also includes information on the proposed methods:

Block diagram, details on the different ML algorithms tested, and a description of processes for evaluation and hyperparameter optimization. Section 3 presents the results and discussion, including the results of hyperparameter optimization and an evaluation of the models with optimal hyperparameters. Section 3 also includes the results of statistical comparisons of different algorithms as well as different features. Section 4 presents the conclusion.

2. Materials and Methods

2.1. Dataset

We studied data from the National Medical and Surgical Center named after N. I. Pirogov of the Russian Healthcare Ministry (Moscow, Russia). The dataset included anonymized results from patients under long-term monitoring from the Department of Neurology and Clinical Neurophysiology in 2017–2019. This procedure aimed to record samples of epileptic activity and verify epileptogenic zones for further clinical treatment. During monitoring, patients kept a regular daily routine with a normal sleep–wake cycle and regular physiological trials. The trials included photic stimulation and hyperventilation, commonly used to stimulate the emergence of epileptiform activity. However, no of the seizures were triggered by these trials, i.e., all epileptic seizures in the resulting dataset were spontaneous. The length of recording varied between patients from 8 to 84 h and depended on the patient's condition. The number of recorded epileptic seizures varied between patients from 1 to 5. All medical procedures followed the Helsinki Declaration and the Center's medical regulations. Patients provided written informed consent before participation.

EEG signals were recorded with a "Micromed" recorder (Micromed S.p.A., Italy). The dataset for each patient consisted of 25 EEG channels with a sampling rate of 128 Hz. During the recording, EEG electrodes were arranged according to a"10–20" scheme with the ground electrode on the forehead and reference electrodes on the ears. EEG signals were examined and labelled by an experienced epileptologist.

Here, we used EEG data of 80 patients diagnosed pathologically with focal epilepsy. However, there was no uniformity of the diagnosis: in different patients, epileptic focus points were found in frontal, temporal, or parietal areas of one or both hemispheres.

2.2. Data Processing

Data processing mainly followed the pipeline described in our previous work [25]. EEG signals are highly susceptible to external and internal noises, and this becomes even more prominent during long-term recording [27]. To increase the quality of EEG data, we applied two steps of pre-processing. Firstly, we filtered EEG data with band-pass (1–60 Hz) and notch (50 Hz) filters. High-pass filter restrained low-frequency artefacts caused by stray effects or breathing, while the low-pass filter reduced high-frequency components, usually related to muscle activity during everyday routine or seizure convulsions [28]. The notch filter suppressed 50 Hz noise from the power grid. Secondly, we used an artefact-removal approach based on independent component analysis (ICA) to mainly remove blinking artefacts.

Further data processing included time–frequency analysis. For this purpose, we used continuous wavelet transform (CWT) with Morlet mother wavelet [29]. We calculated wavelet power (WP) as it is one of the most common CWT-based characteristics for describing the time–frequency structure of EEGs [30]. We chose 1–30 Hz as the frequency band of interest since it is acceptable for studying both normal and pathological EEG activity [31]. CWT can produce a lot of WP data that is difficult to process and use as features for ML methods, so we performed additional steps to reduce data complexity. Firstly, we reduced the spatial dimensionality by averaging the WP over all EEG channels. Secondly, we divided the WP time series into a set of non-overlapping 60 s intervals and calculated the median value inside each interval, hence providing downsampling in the temporal domain. Thus, we obtained WPs averaged over all EEG channels and time intervals and used this data for further feature engineering.

2.3. Feature Engineering

Here, we extended the feature engineering approach from our papers [25,26] by introducing additional features of the EEG signals. Each 60 s interval was considered as an event characterized by a vector of 300 values. These values correspond to the spectrum of WP in the 1–30 Hz band divided into bins with a 0.1 Hz step.

To reduce the complexity of the data, we implemented principal component analysis (PCA) [32]. In PCA, it is important to assess the number of PCs for the output; however, those components must have high explained variance of the original data. Thus, we applied the following algorithm: we decomposed the data into several PCs, introduced a threshold for explained variance—90%, then chose a minimum number of components that could explain a high enough variance. The dependence of the minimum explained variance on the number of PCs is shown in Figure 1. The number of PCs was chosen to be four. Together four PCs could explain at least 98% and 90% of the variance in the original data for the 3 Hz and 30 Hz frequency bands, respectively. In practice, we do not need to use all 300 values of initial WP, but instead can use only four PCs to train the model. This approach simplifies the work of the model, since the number of features used for classification is reduced, while we lose a fairly low percentage of initial information.

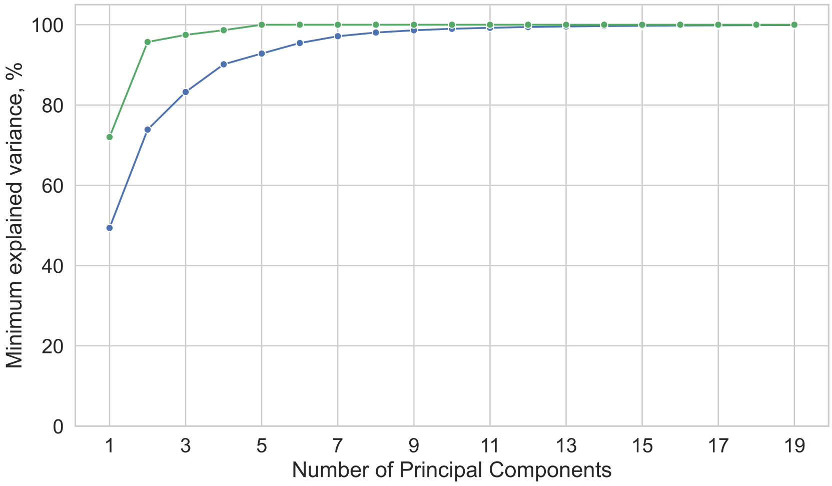

Figure 1. The dependence of the minimum explained variance on the number of PCs for the 3 Hz (green line) and 30 Hz (blue line) frequency bands.

Based on the data, we proposed several features for the ML algorithm, including ones obtained through PCA:

- 30 Hz raw—the vector of all 300 values of the WP in the 1–30 Hz band;
- 3 Hz raw—the sub-vector of 30 values of the WP in the 1–3 Hz band;
- 30 Hz mean—the mean value of the WP in the 1–30 Hz band;
- 3 Hz mean—the mean value of the WP in the 1–3 Hz band;
- 30 Hz PCA—PCA-based feature obtained after decomposing the WP in the 1–30 Hz band and considering four PCs that explain 90% of the variance;
- 3 Hz PCA—PCA-based feature obtained after decomposing the WP in the 1–3 Hz band and considering four PCs that explain 98% of the variance.

We explain chosen features with the following reasoning. It is a well-known fact that epileptic seizures represent abnormal excessive neuronal activity in the brain [2]. Abnormality suggests that EEG signals during seizure episodes differ greatly from normal EEG signals. Hence, the properties of the EEG spectrum should reflect these differences as well. Thus, for spectrum properties, we considered:

- the overall distribution of the WP across frequencies reflected by the feature 30 Hz raw;
- the mean value of the WP in the spectrum reflected by the feature 30 Hz mean;
- the WP values at dominant frequencies reflected by the feature 30 Hz PCA.

Recent works have demonstrated that epileptic activity is associated with characteristic frequency bands in both animal [33] and human subjects [24–26]. For human epilepsy, the band is considered to be 1–3 Hz, so we added three features to reflect spectrum properties in this band: 3 Hz raw, 3 Hz mean, and 3 Hz PCA.

2.4. Machine Learning Methods

We used outlier detection methods, a subgroup of unsupervised ML algorithms. Generally, we considered three different approaches: distance-based methods, SVM, and random forest. Below, we give a brief overview of these methods and describe their main hyperparameters. For a more detailed description, see the recent work of Urs Lenz, Peralta, and Cornelis [34]. Figure 2 shows a block diagram of the method.

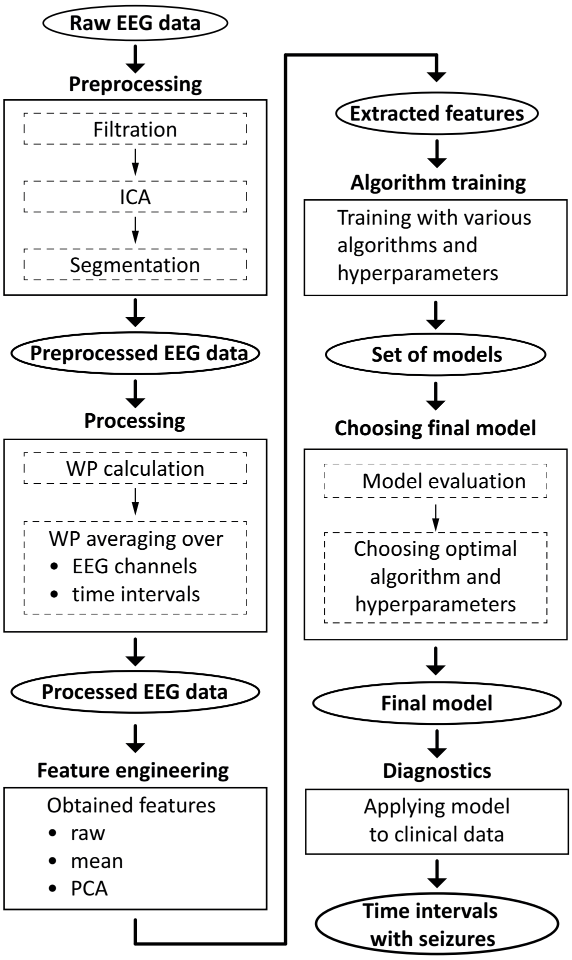

Figure 2. Block diagram of the proposed ML method.

2.4.1. One-Class Support Vector Machine

A one-class support vector machine (OCSVM) was chosen as the first ML method because it has already demonstrated a good performance in the analysis of EEG data [22,35]. SVM takes the data to a higher dimensional space and then aims to find the hyperplane that separates the vectors into two classes. OCSVM, on the other hand, learns to separate normal data from abnormal data. This model has four hyperparameters:

- **Kernel type** is similar to one used in standard SVM classifiers;
- Threshold parameter (**nu**) indicates the expected percentage of outliers in the data;
- Kernel coefficient (**gamma**) determines the degree of wrapping of the vectors by the plane;
- Stopping criterion (**tol**) implies that the algorithm stops running when the difference between old and new loss values becomes less than **tol**.

2.4.2. k-Nearest Neighbors

The k-nearest neighbours (kNN) method works on a simple principle—it calculates the distances from each data vector to its k-nearest vectors and compares these distances to some predefined threshold [36]. kNN has three hyperparameters:

- **Algorithm** is a parameter responsible for the method used for distance calculation;
- **n_neighbors** defines the number of nearest neighbors;
- **Threshold** defines a decision boundary, i.e., the data with a distance exceeding the threshold is referred to as an outlier.

2.4.3. Local Nearest Neighbors Distance

The local nearest neighbours distance (LNND) [37] relies on the same principle as kNN: once the distance to the data exceeds a predefined threshold, these data are marked as outliers. The main difference is that the LNND distance in the attribute space is not valued equally everywhere, but instead it is compared to the local distance between nearby training instances. For instance, if **A** is a given point and **B** is its kth nearest neighbour, then the localized distance is the distance from **A** to **B**, divided by the distance from **B** to its kth nearest neighbour. Hyperparameters for LNND are the same as the ones for kNN.

2.4.4. Local Outlier Factor

The local outlier factor (LOF) method is based on the concept of local density, which reflects how close the neighbours are to a given point [38]. Based on the distances to the nearest neighbours, the algorithm calculates the local density for each data vector and then extracts the vectors whose density is lower than their neighbour's. This method has three hyperparameters:

- **Algorithm** defines a distance measure;
- **n_neighbors** is the number of neighbours;
- **Contamination** sets the percentage of outliers in the dataset.

2.4.5. Isolation Forest

Isolation forest (IF) is an adaptation of the random forest classifier for one-class classification. Its central idea is that instances that are more isolated from the target class should be easier to separate from the training instances. This idea is transferred to the model by constructing randomized search trees on the target data and measuring the average number of steps required to pass through these trees. IF has one hyperparameter—**contamination** [39].

2.5. Evaluation and Hyperparameter Optimization

For ML algorithms, we adjusted hyperparameters to maximize the F1-score [40]. Due to EEG variety, we trained an algorithm on the data of a single patient and calculated the corresponding $F1$-score using the equation:

$$F1 = \frac{2 \times precision \times recall}{precision + recall} \times 100\%, \quad (1)$$

where precision and recall are calculated based on the number of true (TP) and false (FP) positives, and false negatives (FN) as follows:

$$precision = \frac{TP}{TP + FP} \times 100\%, \quad (2)$$

$$recall = \frac{TP}{TP + FN} \times 100\%, \quad (3)$$

Then, we averaged individual $F1$-scores across participants and chose hyperparameters that minimized this averaged $F1$-score. To choose parameters, we used the GridSearch function in Python. GridSearch is a simple function for hyperparameter tuning that sorts through all possible combinations of hyperparameters. It trains models with every possible combination and selects one model with the best metric result given as a target parameter.

Of course, such an approach is much more time-consuming than Bayesian optimization, since its computational cost is calculated as

$$O = \prod_{i=1}^{n} l_i, \tag{4}$$

where l_i is the length of the array of values for the i-th hyperparameter. However with GridSearch we can be fully confident that every combination of hyperparameters has been considered and we have received the best variant. In our work, the GridSearch function aimed to maximize the $F1$-score. Parameters were marked as optimal if the model showed the best result averaged in a group of patients. The range of values after hyperparameter tuning by the GridSearch function are shown in Table 1.

Table 1. The range hyperparameter values examined by the GridSearch method.

Algorithm	Hyperparameter	Range of Values
OCSVM	Nu	10^i, $i \in [-6, -1]$
	Gamma	10^i, $i \in [-6, -1]$, and 'scale'
	Tol	10^i, $i \in [-6, -1]$
	Kernel type	'rbf', 'poly', 'sigmoid'
kNN, LNND, LOF	N_neighbors	$i \in [1, 20]$
	Algorithm	'Euclidean', 'manhattan', 'cosine'
	Threshold (for kNN, LNND), %	$j \times 10^i$, $i \in [-4, 1]$, $j \in 1, 5$
	Contamination (For LOF)	$j \times 10^i$, $i \in [-6, -1]$, $j \in 1, 5$
IF	Contamination	$j \times 10^i$, $i \in [-6, -1]$, $j \in 1, 5$

Finally, to generate precision–recall curves, we calculated the probabilities for each event to be classified as a seizure and varied the decision boundary.

We tested the differences in $F1$-score between the algorithms using the Friedman test. For the post hoc analysis, we used the non-parametric Conover's test [41].

For performance evaluation, we used the criteria most suitable for the considered epileptic EEG data characterized by a high level of class imbalance. Accuracy and specificity would always be high since only 2% of the data would be positive, so these metrics would not reflect the real quality of the ML models. We chose sensitivity (recall) and precision, as these two metrics reflect the ability of the model to search the seizures. These metrics are shown in the form of precision–recall curves. Since GridSearch can only focus on one metric, we combined precision and recall into the $F1$-score, helping to properly assess the quality of the trained models.

3. Results and Discussion

We started with hyperparameter optimization for all algorithms and all types of input data (Table 2). We tested the models with different types of distance estimation metrics, and the GridSearch method chose the Euclidean distance as the best metric.

Therefore, we proceeded with the Euclidean distance in all calculations. The OCSVM performed better with the rbf kernel when the gamma was set to scale. The obtained results provided the following insights. First, the LNND and LOF required more neighbours than kNN. Second, the parameters of the OCSVM did not depend on the type of input data. Third, in the distance-based methods the threshold had the smallest value for PCA and the highest value for the 30 Hz and 3 Hz spectral power.

Table 2. The optimal hyperparameters that provide the highest $F1$-score for the different algorithms and types of input data.

Algorithm	**Hyperparameter**	**Input Data**					
		30 Hz	**3 Hz**	**30 Hz Mean**	**3 Hz Mean**	**30 Hz PCA**	**3 Hz PCA**
OCSVM	Nu	0.1	0.1	0.1	0.1	0.1	0.1
	Gamma	scale					
	Tol	0.1	0.1	0.1	0.1	0.1	0.1
	Kernel type	rbf					
kNN	N_neighbors	3	3	4	4	4	3
	Threshold, %	0.5	0.5	0.1	0.1	0.1	0.1
	Algorithm	Euclidean					
LNND	N_neighbors	5	8	9	9	9	7
	Threshold, %	0.5	0.5	0.5	0.5	0.1	0.1
	Algorithm	Euclidean					
LOF	N_neighbors	7	8	8	8	8	3
	Contamination	0.005	0.005	0.001	0.005	0.001	0.0001
	Algorithm	Euclidean					
IF	Contamination	0.001	0.005	0.001	0.001	0.001	0.005

After finding the optimal combination of hyperparameters, an individual model was trained for each patient, i.e., 80 models were obtained for each ML method. Then, the recall, precision, and $F1$-score were calculated, and these values were averaged for a group of patients (see Table 3). These results show that OCSVM, kNN, and IF reach the highest performance on the raw data, while the LNND and LOF perform better on PCAs. The best results were demonstrated by the LNND trained using 30 Hz PCAs.

Table 3. The maximum $F1$-score for all models trained on the different types of input data using the optimal parameters from Table 2. Results are shown as a group mean and 95% confidence intervals (bold text indicates the best results of the models).

Feature	**Algorithm**				
	OCSVM	**kNN**	**LNND**	**LOF**	**IF**
3 Hz Raw	0.305 ± 0.057	**0.316 ± 0.055**	0.281 ± 0.055	0.297 ± 0.055	**0.304 ± 0.057**
30 Hz Raw	**0.307 ± 0.060**	0.312 ± 0.060	0.331 ± 0.056	0.330 ± 0.054	0.271 ± 0.073
3 Hz mean	0.255 ± 0.071	0.282 ± 0.074	0.116 ± 0.040	0.278 ± 0.073	0.282 ± 0.074
30 Hz mean	0.245 ± 0.071	0.270 ± 0.073	0.135 ± 0.046	0.254 ± 0.053	0.270 ± 0.073
3 Hz PCA	0.273 ± 0.072	0.300 ± 0.075	0.250 ± 0.071	0.292 ± 0.075	0.276 ± 0.073
30 Hz PCA	0.300 ± 0.058	0.313 ± 0.077	**0.338 ± 0.077**	**0.331 ± 0.080**	0.304 ± 0.061

To test if the $F1$-score changed between algorithms, we used the non-parametric Friedman test.

For the mean value in the 3 Hz band, there was a statistically significant difference in the $F1$-score depending on which algorithm was used, $\chi^2(4) = 91.731$, $p < 0.001$. Post hoc analysis with Conover's tests was conducted with a Bonferroni correction applied, resulting in a significance level set at $p < 0.01$. There was a significant difference in the $F1$-score between the LNND and other algorithms (Table 4).

Table 4. 3 Hz mean (Conover's post hoc comparisons).

		T-Stat	df	W_i	W_j	p	p_{bonf}	p_{holm}
OCSVM	kNN	0.039	316	264.000	263.500	0.969	1.000	1.000
	LNND	7.883	316	264.000	163.500	<0.001	<0.001	<0.001
	LOF	1.294	316	264.000	247.500	0.197	1.000	1.000
	IF	0.196	316	264.000	261.500	0.845	1.000	1.000
kNN	LNND	7.844	316	263.500	163.500	<0.001	<0.001	<0.001
	LOF	1.255	316	263.500	247.500	0.210	1.000	1.000
	IF	0.157	316	263.500	261.500	0.875	1.000	1.000
LNND	LOF	6.589	316	163.500	247.500	<0.001	<0.001	<0.001
	IF	7.687	316	163.500	261.500	<0.001	<0.001	<0.001
LOF	IF	1.098	316	247.500	261.500	0.273	1.000	1.000

For the mean value in the 30 Hz band, there was a statistically significant difference in the F1-score depending on which algorithm was used, $\chi^2(4) = 56.151, p < 0.001$. Post hoc analysis with Conover's tests was conducted with a Bonferroni correction applied, resulting in a significance level set at $p < 0.01$. There was a significant difference in the F1-score between the LNND and other algorithms (Table 5).

For the PCA in the 3 Hz band, there was no statistically significant difference in the F1-score between the algorithms, $\chi^2(4) = 6.616, p < 0.158$.

For the PCA in the 30 Hz band, there was no statistically significant difference in the F1-score between the algorithms, $\chi^2(4) = 0.988, p < 0.912$.

For the RAW data in the 3 Hz band, there was no statistically significant difference in the F1-score between the algorithms, $\chi^2(4) = 4.934, p < 0.212$.

For the RAW data in the 30 Hz band, there was no statistically significant difference in the F1-score between the algorithms, $\chi^2(4) = 3.253, p < 0.516$.

Table 5. 30 Hz mean (Conover's post hoc comparisons).

		T-Stat	df	W_i	W_j	p	p_{bonf}	p_{holm}
OCSVM	kNN	1.013	316	244.000	254.000	0.312	1.000	1.000
	LNND	5.115	316	244.000	193.500	<0.001	<0.001	<0.001
	LOF	0.861	316	244.000	252.500	0.390	1.000	1.000
	IF	1.215	316	244.000	256.000	0.225	1.000	1.000
kNN	LNND	6.128	316	254.000	193.500	<0.001	<0.001	<0.001
	LOF	0.152	316	254.000	252.500	0.879	1.000	1.000
	IF	0.203	316	254.000	256.000	0.840	1.000	1.000
LNND	LOF	5.976	316	193.500	252.500	<0.001	<0.001	<0.001
	IF	6.331	316	193.500	256.000	<0.001	<0.001	<0.001
LOF	IF	0.355	316	252.500	256.000	0.723	1.000	1.000

The $F1$-scores appeared to be low; however, there are certain reasons for this. Firstly, in unsupervised ML, we cannot compare predicted values to true values and change the parameters of the method to improve the performance. Since unsupervised methods aim to separate anomalies from the rest of the data, the main way to affect performance is to estimate the percentage of outliers in the data. Secondly, EEG data contains a lot of artefacts that can also be marked as anomalies, and since the data is not pre-labelled, we have limited options to avoid them. The presence of artefacts affects the threshold value for the method, so in highly contaminated data, the $F1$-score can be lower. Nonetheless, even these F1-scores can be beneficial for DSS, where the final decision is made by a human. The proposed system could be used for express analyses of the EEG to save the time and effort of experts.

In the next step, we considered the precision–recall curves for all algorithms trained using the optimal hyperparameters from Table 2 on the different types of input data (Figure 3). Each sub-figure shows the precision–recall curves for all algorithms (different colours).

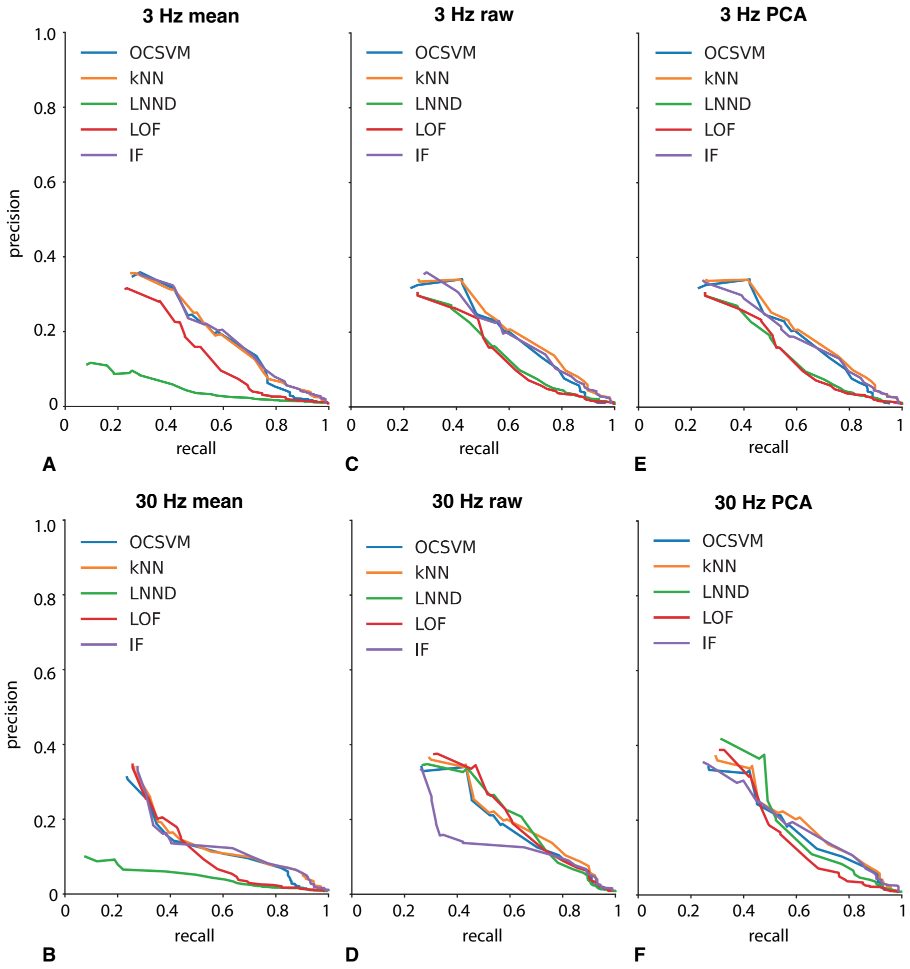

Figure 3. Precision–recall curves for the different outlier detection algorithms are shown in different colours. Each panel reflects the type of input data as stated in the panel caption: (**A**) 3 Hz mean; (**B**) 30 Hz mean; (**C**) 3 Hz raw; (**D**) 30 Hz raw; (**E**) 3 Hz PCA; (**F**) 30 Hz PCA.

The obtained results provided the following insights. First, the lowest areas under the precision–recall curves were achieved at 30 Hz mean (Figure 3B). This indicates that the overall power may not be a reliable marker of a seizure. In particular, the precision is low, indicating a large number of false detections. The possible explanation is that there are many events characterized by an increase in overall power, such as artefacts or sleep-related patterns [42]. Second, using 3 Hz mean improves the precision–recall curves (Figure 3A). This illustrates that seizures produce an outlier in a specific frequency band rather than increasing the overall power. In line with our previous results, this confirms that extreme behaviour occurs in a certain frequency range [24,33]. Third, the mean spectral power provides the worst results for the LOF and LNND algorithms (red and green curves). Moreover, the transition to the specific frequency band of 3 Hz barely improves the performance of these methods. We see that the green and red curves practically do not change between Figure 3A,B, while the others show higher precision. Note that LOF and LNND are distance-based methods, working similarly to kNN but using a different distance measure. Knowing that kNN improves performance when switching from 30 Hz mean to 3 Hz mean, we hypothesize that the distance measure affects the separability between the classes. Fourth, using the entire 30 Hz range provides the widest feature space in which we see that the LOF and LNND perform better. The possible reason for this is that the boundary between the classes changes its configuration in the different feature spaces. Fifth, the best precision–recall curve is achieved using the LNND trained on 30 Hz PCA (Figure 3F). Interestingly, the LNND improves its performance when using PCAs in the 30 Hz range rather than 3 Hz. An intuitively clear explanation is that extending the frequency range provides more distinguishing features together with an increasing amount of unnecessary information and correlated features, therefore limiting the ability to improve. Using PCAs reduces the amount of unnecessary information and rejects correlations between features but leaves the advantage of using a wide frequency range.

Finally, we analysed the effect of the distance measure in the distance-based algorithms on the separability between seizures and normal activity in the feature space. We introduced distance between the classes, a normalized difference between the median distances to seizures and normal states, and tested if it was dependent on the algorithm, frequency band, or feature (see Figure 4). The median distances did not follow a normal distribution in the group of patients (according to the Shapiro–Wilk test), having long tails over the high values. Therefore, we removed 5% of patients with the highest values and transformed the rest data using the root-mean-square. Resulting values underwent repeated measures of analysis of variances (rm ANOVA) with two within-subject factors: frequency band (3 Hz and 30 Hz) and feature (raw, mean, PCA) separately for three ML algorithms (LNND, kNN, and LOF).

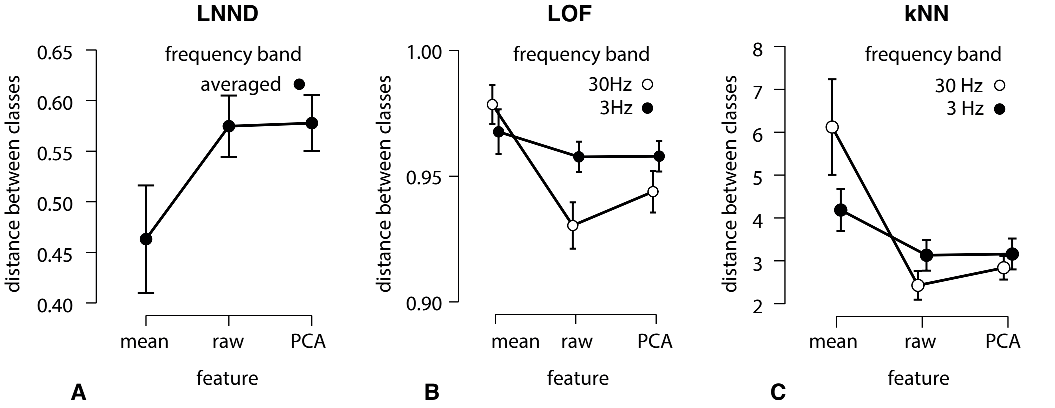

Figure 4. Distance between the seizures and normal states (group mean and 95% CI) in the feature space depending on the frequency band and feature. Sub-figures correspond to the different distance-based ML algorithms: LNND (**A**); LOF (**B**); kNN (**C**).

For the LNND, the rm ANOVA reveals a significant main effect of the feature: F(2,106) = 11.468, $p < 0.001$. The main effect of the frequency band was insignificant: F(1,53) = 1.472, $p = 0.23$. Finally, the interaction effect feature × frequency band was also insignificant: F(2,106) = 1.365, $p < 0.26$. These results evidence that the distance between the classes depends on the feature in a similar way for both frequency bands. The direction of change is shown in Figure 4A. Note that since the effect of frequency and the feature × frequency band interaction were both insignificant, the distance between classes in Figure 4A is presented as an average between the 30 Hz and 3 Hz bands. Using the mean value as a feature provides the smallest distance between the seizures and normal states. Raw and PCA ensure a greater distance but it barely differs between them.

For the LOF, the rm ANOVA reveals a significant main effect of the feature: F(2,78) = 53.738, $p < 0.001$, and an insignificant main effect of the frequency band: F(1,39) = 3.757, $p = 0.06$. In contrast, there was a significant feature × frequency band interaction effect: F(2,78) = 23.109, $p < 0.001$. These results evidence that the distance between the classes depends on the feature and the way it changes depends on the frequency band. The direction of change is shown in Figure 4B for 30 Hz and 3 Hz. For the 3 Hz band, the distance between the classes hardly changes between the different features. For the

30 Hz band, the mean value provides the highest distance, and raw data provides the lowest distance.

For the kNN, the rm ANOVA reveals a significant main effect of the feature: F(2,114) = 35.854, $p < 0.001$, and an insignificant main effect of the frequency band: F(1,57) = 1.312, $p = 0.257$. There was a significant feature × frequency band interaction effect: F(2,114) = 19.894, $p < 0.001$. These results evidence that the distance between the classes depends on the feature and the way it changes depends on the frequency band. The direction of change is shown in Figure 4C for 30 Hz and 3 Hz. When using the mean as a feature, the distance differs between 3 Hz and 30 Hz, for the raw and PCA this difference disappears.

Combining the results of outlier detection (Figure 3) and the analysis of distances (Figure 4), we find similar tendencies in the dependence of the precision–recall curve and distance in the LNND on the frequency band and feature. First, the lowest precision–recall curve is observed for the mean value and grows when we use raw data and PCA. Second, this tendency remains similar for 3 Hz and 30 Hz. Therefore, we conclude that in the LNND, the structure of the feature space affects the distance between the classes (i.e., separability). Estimating the separability may give prior knowledge about the performance of the LNND, therefore enabling the selection of features to ensure the best performance.

In terms of achievable performance rates, the obtained results are consistent with modern studies, which also propose novel methods for automated seizure identification in EEG signals using different supervised ML-based approaches. Amiri et al., proposed [9] a method utilizing sparse common spatial patterns (sCSPs) and adaptive short-time Fourier transform-based synchrosqueezing transforms (adaptive FSSTs) to enhance the time–frequency representation of multi-component EEG signals and reduce noise and interferences. They method achieved outstanding performance with high sensitivity, specificity, and accuracy in detecting seizures, demonstrating the potential of their proposed method for epilepsy diagnosis. Malekzadeh et al., utilized [10] a computer-aided diagnosis system (CADS) for the automatic diagnosis of epileptic seizures in EEG signals, involving pre-processing, feature extraction, and classification steps. They used tunable-Q wavelet transforms (TQWTs) for EEG signal decomposition and extracted various linear and non-linear features from TQWT sub-bands. They employed different approaches based on conventional ML and DL for classification and achieved high accuracy (up to 99%) in detecting seizures using their proposed DL method based on convolutional neural network (CNN) and RNN. Jiwani et al., proposed [11] an LSTM-CNN model for epileptic seizure detection using EEG signals, highlighting the need for automated techniques due to the lengthy process and shortage of specialists for the visual inspection of EEG reports. Their proposed model combined LSTM and CNN for feature extraction and classification, and they achieved promising results in detecting seizures.

Thus, our findings provide valuable insights into the performance of outlier detection algorithms for epileptic seizure detection in EEG signals and highlight the potential of distance-based methods, particularly when combined with PCA, for achieving high performance in this task. These findings are consistent with current studies [9–12]. Further research in this area could potentially lead to the development of more effective and efficient computer-aided tools for epilepsy diagnosis, reducing the interpretation load on specialists and improving patient care.

4. Conclusions

We tested the performance of popular outlier detection algorithms in labelling epileptic seizures in human EEGs. We showed that distance-based methods may outperform SVMs and random forests if the distance calculation and feature space are properly defined. Thus, we obtained the best score from the LNND trained on four PCAs explaining 90% of the variance in the spectral power in the 1–30 Hz frequency range. Finally, we tested if the distance between the classes (separability) in feature space, a core of the distance-based algorithm, predicts its performance. We observed that PCA provided better separability regardless of the frequency band. Therefore, we suggest that estimating the distance between the classes may predict the algorithm's performance providing opportunities for feature selection.

Author Contributions: Conceptualization, O.E.K., D.A.A. and A.E.H.; data curation, O.E.K., V.V.G. and N.U.; formal analysis, V.A.M.; funding acquisition, D.A.A. and A.E.H.; investigation, M.S.K.; methodology, O.E.K., S.A.K., V.A.M. and A.E.H.; project administration, O.E.K. and D.A.A.; resources, O.E.K., N.U. and D.A.A.; software, M.S.K. and V.V.G.; supervision, A.E.H.; validation, V.A.M., V.V.G., S.A.K., M.S.K. and A.E.H.; visualization, S.A.K. and M.S.K.; writing—original draft, V.A.M. and V.V.G.; writing—review and editing, O.E.K., S.A.K. and A.E.H. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Academic Leadership Program "PRIORITY-2030" of Immanuel Kant Baltic Federal University of Ministry of Science and Education of Russian Federation. AEH thanks the Leading Scientific School Support Program (grant NSH-589.2022.1.2). SAK thanks the Doctor Support Program (grant MD-590.2022.1.2).

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Ethics Committee of Immanuel Kant Baltic Federal University (protocol code 7/2022 of 27 June 2022).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data that support the findings of this study are available from the National Medical and Surgical Center named after N. I. Pirogov of Russian Healthcare Ministry but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of the National Medical and Surgical Center named after N. I. Pirogov of Russian Healthcare Ministry.

Conflicts of Interest: The authors declare no conflicts of interest. The funders and organizations had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

EEG	Electroencephalogram
ML	Machine learning
DL	Deep learning
RNN	Recurrent neural networks
LSTM	Long short-term memory
DSS	Decision support system
SVM	Support vector machine
PC	Principle component
ICA	Independent component analysis
CWT	Continuous wavelet transform
WP	Wavelet power
PCA	Principle component analysis
OCSVM	One-class support vector machine
kNN	k-Nearest neighbours
LNND	Local nearest neighbours distance
LOF	Local outlier factor
IF	Isolation forest
TP	True positive
FP	False positive
FN	False negative
rm ANOVA	Repeated measures analysis of variances
sCSP	Sparse common spatial pattern
FSST	Fourier transform-based synchrosqueezing transform

CADS	Computer-aided diagnosis system
TQWT	Tunable-Q wavelet transform
CNN	Convolutional neural network

References

1. Beghi, E. The epidemiology of epilepsy. *Neuroepidemiology* **2020**, *54*, 185–191. [CrossRef]
2. Thijs, R.D.; Surges, R.; O'Brien, T.J.; Sander, J.W. Epilepsy in adults. *Lancet* **2019**, *393*, 689–701. [CrossRef] [PubMed]
3. Fisher, R.S.; Acevedo, C.; Arzimanoglou, A.; Bogacz, A.; Cross, J.H.; Elger, C.E.; Engel, J., Jr.; Forsgren, L.; French, J.A.; Glynn, M.; et al. ILAE official report: A practical clinical definition of epilepsy. *Epilepsia* **2014**, *55*, 475–482. [CrossRef] [PubMed]
4. Goldberg, E.M.; Coulter, D.A. Mechanisms of epileptogenesis: A convergence on neural circuit dysfunction. *Nat. Rev. Neurosci.* **2013**, *14*, 337–349. [CrossRef] [PubMed]
5. Motamedi, G.; Meador, K. Epilepsy and cognition. *Epilepsy Behav.* **2003**, *4*, 25–38. [CrossRef] [PubMed]
6. Elger, C.E.; Hoppe, C. Diagnostic challenges in epilepsy: Seizure under-reporting and seizure detection. *Lancet Neurol.* **2018**, *17*, 279–288. [CrossRef]
7. Friedman, D.E.; Hirsch, L.J. How long does it take to make an accurate diagnosis in an epilepsy monitoring unit? *J. Clin. Neurophysiol.* **2009**, *26*, 213–217. [CrossRef]
8. Tatum, W.O. *Handbook of EEG Interpretation*; Springer Publishing Company: Berlin/Heidelberg, Germany, 2021.
9. Amiri, M.; Aghaeinia, H.; Amindavar, H.R. Automatic epileptic seizure detection in EEG signals using sparse common spatial pattern and adaptive short-time Fourier transform-based synchrosqueezing transform. *Biomed. Signal Process. Control* **2023**, *79*, 104022. [CrossRef]
10. Malekzadeh, A.; Zare, A.; Yaghoobi, M.; Kobravi, H.R.; Alizadehsani, R. Epileptic seizures detection in EEG signals using fusion handcrafted and deep learning features. *Sensors* **2021**, *21*, 7710. [CrossRef]
11. Jiwani, N.; Gupta, K.; Sharif, M.H.U.; Adhikari, N.; Afreen, N. A LSTM-CNN Model for Epileptic Seizures Detection using EEG Signal. In Proceedings of the 2022 2nd International Conference on Emerging Smart Technologies and Applications (eSmarTA), Ibb, Yemen, 25–26 October 2022; pp. 1–5.
12. Khan, I.M.; Khan, M.M.; Farooq, O. Epileptic Seizure Detection using EEG Signals. In Proceedings of the 2022 5th International Conference on Computing and Informatics (ICCI), New Cairo, Egypt, 9–10 March 2022; pp. 111–117.
13. Siddiqui, M.K.; Morales-Menendez, R.; Huang, X.; Hussain, N. A review of epileptic seizure detection using machine learning classifiers. *Brain Inform.* **2020**, *7*, 5. [CrossRef]
14. Mohri, M.; Rostamizadeh, A.; Talwalkar, A. *Foundations of Machine Learning*; MIT Press: Cambridge, MA, USA, 2018.
15. Tzimourta, K.D.; Tzallas, A.T.; Giannakeas, N.; Astrakas, L.G.; Tsalikakis, D.G.; Angelidis, P.; Tsipouras, M.G. A robust methodology for classification of epileptic seizures in EEG signals. *Health Technol.* **2019**, *9*, 135–142. [CrossRef]
16. Ullah, I.; Hussain, M.; Qazi, E.-u.-H.; Aboalsamh, H. An automated system for epilepsy detection using EEG brain signals based on deep learning approach. *Expert Syst. Appl.* **2018**, *107*, 61–71. [CrossRef]
17. Pominova, M.; Artemov, A.; Sharaev, M.; Kondrateva, E.; Bernstein, A.; Burnaev, E. Voxelwise 3D convolutional and recurrent neural networks for epilepsy and depression diagnostics from structural and functional MRI data. In Proceedings of the 2018 IEEE International Conference on Data Mining Workshops (ICDMW), Singapore, 17–20 November 2018; pp. 299–307.
18. Abdelhameed, A.M.; Daoud, H.G.; Bayoumi, M. Deep convolutional bidirectional LSTM recurrent neural network for epileptic seizure detection. In Proceedings of the 2018 16th IEEE International New Circuits and Systems Conference (NEWCAS), Montreal, QC, Canada, 24–27 June 2018; pp. 139–143.
19. Si, Y. Machine learning applications for electroencephalograph signals in epilepsy: A quick review. *Acta Epileptol.* **2020**, *2*, 5. [CrossRef]
20. Birjandtalab, J.; Pouyan, M.B.; Nourani, M. Unsupervised eeg analysis for automated epileptic seizure detection. In Proceedings of the First International Workshop on Pattern Recognition, Tokyo, Japan, 11–13 May 2016; International Society for Optics and Photonics: Bellingham, WA, USA, 2016; Volume 10011, p. 100110M.
21. Wickramasinghe, C.S.; Amarasinghe, K.; Marino, D.L.; Rieger, C.; Manic, M. Explainable unsupervised machine learning for cyber-physical systems. *IEEE Access* **2021**, *9*, 131824–131843. [CrossRef]
22. Chen, Z.; Lu, G.; Xie, Z.; Shang, W. A unified framework and method for EEG-based early epileptic seizure detection and epilepsy diagnosis. *IEEE Access* **2020**, *8*, 20080–20092. [CrossRef]
23. Nandan, M.; Talathi, S.S.; Myers, S.; Ditto, W.L.; Khargonekar, P.P.; Carney, P.R. Support vector machines for seizure detection in an animal model of chronic epilepsy. *J. Neural Eng.* **2010**, *7*, 036001. [CrossRef] [PubMed]
24. Karpov, O.E.; Grubov, V.V.; Maksimenko, V.A.; Utaschev, N.; Semerikov, V.E.; Andrikov, D.A.; Hramov, A.E. Noise amplification precedes extreme epileptic events on human EEG. *Phys. Rev. E* **2021**, *103*, 022310. [CrossRef]
25. Karpov, O.E.; Grubov, V.V.; Maksimenko, V.A.; Kurkin, S.A.; Smirnov, N.M.; Utyashev, N.P.; Andrikov, D.A.; Shusharina, N.N.; Hramov, A.E. Extreme value theory inspires explainable machine learning approach for seizure detection. *Sci. Rep.* **2022**, *12*, 11474. [CrossRef]
26. Karpov, O.E.; Afinogenov, S.; Grubov, V.V.; Maksimenko, V.; Korchagin, S.; Utyashev, N.; Hramov, A.E. Detecting epileptic seizures using machine learning and interpretable features of human EEG. *Eur. Phys. J. Spec. Top.* **2022**, 1–10. [CrossRef]

27. White, D.M.; Van Cott, C.A. EEG artifacts in the intensive care unit setting. *Am. J. Electroneurodiagn. Technol.* **2010**, *50*, 8–25. [CrossRef]
28. Ebersole, J.S.; Pedley, T.A. *Current Practice of Clinical Electroencephalography*; Lippincott Williams & Wilkins: Pennsylvania Furnace, PA, USA, 2003.
29. Aldroubi, A.; Unser, M. *Wavelets in Medicine and Biology*; Routledge: Oxfordshire, UK, 2017.
30. Hramov, A.E.; Koronovskii, A.A.; Makarov, V.A.; Maximenko, V.A.; Pavlov, A.N.; Sitnikova, E. *Wavelets in Neuroscience*; Springer: Berlin/Heidelberg, Germany, 2021.
31. Adeli, H.; Zhou, Z.; Dadmehr, N. Analysis of EEG records in an epileptic patient using wavelet transform. *J. Neurosci. Methods* **2003**, *123*, 69–87. [CrossRef] [PubMed]
32. Bro, R.; Smilde, A.K. Principal component analysis. *Anal. Methods* **2014**, *6*, 2812–2831. [CrossRef]
33. Frolov, N.S.; Grubov, V.V.; Maksimenko, V.A.; Lüttjohann, A.; Makarov, V.V.; Pavlov, A.N.; Sitnikova, E.; Pisarchik, A.N.; Kurths, J.; Hramov, A.E. Statistical properties and predictability of extreme epileptic events. *Sci. Rep.* **2019**, *9*, 7243. [CrossRef] [PubMed]
34. Lenz, O.U.; Peralta, D.; Cornelis, C. Average Localised Proximity: A new data descriptor with good default one-class classification performance. *Pattern Recognit.* **2021**, *118*, 107991. [CrossRef]
35. Burnaev, E.; Smolyakov, D. One-class SVM with privileged information and its application to malware detection. In Proceedings of the 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW), Barcelona, Spain, 12–15 December 2016; pp. 273–280.
36. Zhang, Z. Introduction to machine learning: k-nearest neighbors. *Ann. Transl. Med.* **2016**, *4*, 218. [CrossRef] [PubMed]
37. Zheng, W.; Zhao, L.; Zou, C. Locally nearest neighbor classifiers for pattern classification. *Pattern Recognit.* **2004**, *37*, 1307–1309. [CrossRef]
38. Breunig, M.M.; Kriegel, H.P.; Ng, R.T.; Sander, J. LOF: Identifying density-based local outliers. In Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, Dallas, TX, USA, 16–18 May 2000; pp. 93–104.
39. Liu, F.T.; Ting, K.M.; Zhou, Z.H. Isolation forest. In Proceedings of the 2008 Eighth IEEE International Conference on Data Mining, Pisa, Italy, 15–19 December 2008; pp. 413–422.
40. Goutte, C.; Gaussier, E. A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. In Proceedings of the Advances in Information Retrieval: 27th European Conference on IR Research (ECIR 2005), Santiago de Compostela, Spain, 21–23 March 2005; pp. 345–359.
41. Conover, W.J. *Practical Nonparametric Statistics*; John Wiley & Sons: Hoboken, NJ, USA, 1999; Volume 350.
42. Maksimenko, V.A.; Van Heukelum, S.; Makarov, V.V.; Kelderhuis, J.; Lüttjohann, A.; Koronovskii, A.A.; Hramov, A.E.; Van Luijtelaar, G. Absence seizure control by a brain computer interface. *Sci. Rep.* **2017**, *7*, 2487. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.