MEASUREMENT SCIENCE REVIEW

Identification of Microrecording Artifacts with Wavelet Analysis and Convolutional Neural Network: An Image Recognition Approach

Ondřej Klempíř[1], Radim Krupička[1], Eduard Bakštein[2,3], Robert Jech[4]

[1]*Department of Biomedical Informatics, Faculty of Biomedical Engineering, Czech Technical University in Prague, nám. Sítná, 3105, 272 01, Kladno, Czech Republic, klempond@gmail.com*
[2]*Department of Cybernetics, Faculty of Electrical Engineering, Czech Technical University in Prague, Karlovo náměstí, 13, 121 35, Prague, Czech Republic*
[3]*National Institute of Mental Health, Topolová, 748, 250 67, Klecany, Czech Republic*
[4]*Department of Neurology and Center of Clinical Neuroscience, First Faculty of Medicine and General University Hospital, Charles University, Kateřinská, 468/30, 120 00, Prague, Czech Republic*

Deep brain stimulation (DBS) is an internationally accepted form of treatment option for selected patients with Parkinson's disease and dystonia. Intraoperative extracellular microelectrode recordings (MER) are considered as the standard electrophysiological method for the precise positioning of the DBS electrode into the target brain structure. Pre-processing of MERs is a key phase in clinical analysis, with intraoperative microelectrode recordings being prone to several artifact groups (up to 25 %). The aim of this methodological article is to provide a convolutional neural network (CNN) processing pipeline for the detection of artifacts in an MER. We applied continuous wavelet transform (CWT) to generate an over-complete time–frequency representation. We demonstrated that when attempting to find artifacts in an MER, the new CNN + CWT provides a high level of accuracy (ACC = 88.1 %), identifies individual classes of artifacts (ACC = 75.3 %) and also offers artifact time onset detail, which can lead to a reduction in false positives/negatives. In summary, the presented methodology is capable of identifying and removing various artifacts in a comprehensive database of MER and represents a substantial improvement over the existing methodology. We believe that this approach will assist in the proposal of interesting clinical hypotheses and will have neurologically relevant effects.

Keywords: Artifacts detection, convolutional neural networks, Parkinson's disease, deep brain stimulation, microrecording, wavelet analysis.

1. INTRODUCTION

Major breakthroughs in the treatment of many neurological disorders have been made possible with miniature specialized electronic hardware devices that are implanted in the patient to compensate for impaired physiological functions [1]. In addition to pacemakers and cochlear implants, it is necessary to utilize electrodes used for deep brain stimulation (DBS), which is an effective alternative to the pharmacological treatment of Parkinson's disease (PD) or heterogeneous dystonia [2]. The process of DBS implantation is demanding in terms of time, expertise, and finances. This invasive method is capable of significant symptom alleviation, which could dramatically improve the quality of life of patients with various movement disorders.

The reliable electrophysiological technique to determine accurate DBS electrode placement is intraoperative extracellular microrecording (mEEG; MER) exploration.

Despite advances in magnetic resonance imaging (MRI) technology and a better understanding of the relationship between an MRI and MER, MER evaluation remains the most widely used localization technique [3], [4]. The MER provides a high temporal resolution (< 1 ms) of the brain's neural activity in the close vicinity to the microelectrode [5].

The potential of the development of beneficial medical diagnostic computer programs is enhanced by machine learning methods. In general, the application of deep learning in recent years is increasing and has been shown to be a more accurate and progressive technique in various fields of medical applications, e.g., in computer vision, natural language processing, electronic health record data or bioinformatics [6], [7]. In particular, many studies that deal with the analysis of biomedical signals can be documented.

Bursa et al. applied a continuous wavelet transform (CWT) to a fetal heart rate signal processing with different

levels of time/frequency detail parameters and in two different resolutions. The output 2D structures are fed to a convolutional neural network (CNN) [8]. A CNN is one of the main design representatives of deep learning, which is inspired by the organization of the animal visual cortex. Many challenges for object classification, detection or semantic segmentation have been solved in the field of neuroscience [9]. The findings described in a paper titled CNN to Model Articulation Impairments in Patients with PD, presented new results in speech impairments using deep learning algorithms, the Fourier transform and a CWT deployed on a voiced and unvoiced segment for the classification of PD vs. healthy speakers [10]. Another clinical study demonstrated the feasibility of using deep learning as part of an electromyographic hand gesture signal classification system [11]. Furthermore, Acharya et al. designed and predicted normal, preictal, and seizure EEG signals, based on a 13-layer CNN [12].

Up-to-date research has uncovered findings on the many applications of deep learning in biosignal event detection, specifically in the field of cardiology [13]. For example, Yildirim et al. proposed an algorithm for arrhythmia detection [14]. Researchers found that the AI algorithm successfully predicted distinct diagnostic classes encompassing a normal sinus rhythm, pacemaker rhythm, and other rhythm disorders effectively from an analysis of 10-s ECG signal fragments. In addition to the CNN, a recurrent long-short term memory network is suitable for an ECG classification [15]. However, there have not been any studies to address the potential benefits of MER-DBS processing via deep learning.

Currently, there is a wide variety of powerful deep learning frameworks including, e.g., MATLAB for Deep Learning [16], TensorFlow [17] with Keras [18], PyTorch [19], and Caffe2 [20]. Most of them support interoperability with open source deep learning frameworks using ONNX import and export capabilities.

Intraoperative microelectrode recordings are prone to several artifacts (up to 25 %). Essentially, the three most common classes of artifacts that can be identified [21] are:

• A mechanical movement artifact, manifested by short-time, high-power signal peaks, usually spread across the whole frequency spectrum.

• Low-frequency interference below the mains frequency (50 Hz), causing visible variation in the signal offset or in the baseline.

• Electromagnetic interference at one or multiple stable frequencies, well localized in a narrow band(s) in the frequency spectrum and stable over time. The frequency of the observed long-term interference often differs from the expected odd harmonics of the mains frequency (50 Hz, 150 Hz, 250 Hz, etc.).

Detection of these artifacts is not straightforward, potential artifacts emerge from several sources and may not be present in all signals. Detection of artifacts is an essential phase of MER processing and can have a significant influence on spike sorting [21], and subsequently the determination of clinically interesting biomarkers. Therefore, prior to spike detection, artifacts should

automatically be removed from the signal in order to prevent spike-dependent statistics (i.e., compound firing rate) and spectral measures from being biased [22]. In clinical practice, sensors containing the artifacts are typically manually curated and artifacts are removed after visual inspection.

Artifact detection in MER signals is an area with few objectively validated methods [21], [23]-[26] acting as change-point detection or detectors of significant changes compared to the clean signal spectrum. Many researchers use their own (semi)automatic methods, ranging from simple amplitude thresholding [27] through statistical testing of the amplitude distribution in short signal windows [28].

A distinction has to be made between the externally induced artifacts (i.e., the scope of this paper) and "background noise" in MER, which commonly refers to the recorded activity of neurons further away from the electrode in single/multiunit activity processing [29].

The main goal of this article is to develop an algorithm capable of detecting individual classes of artifacts, corresponding artifacts onset times and localization of the area in the MER in which artifacts originate. In this paper, we present a novel CNN approach to microelectrode activity artifact detection by means of neural networks in biomeasurement. Fig.1. schematically shows the network architecture and the main task presented in this article.
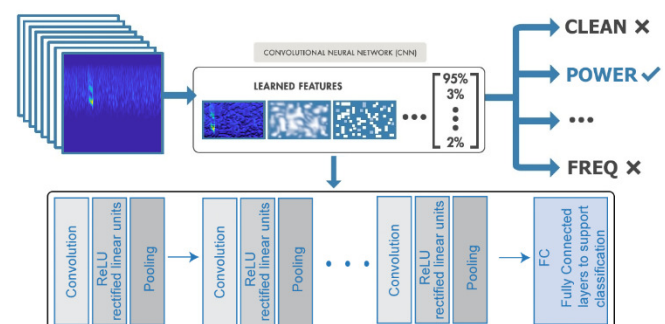

Fig.1. A deep learning workflow. For example, images are passed to the CNN, which automatically learns features and classifies MER segments. Conceptual illustration adapted with quoting [30].

2. SUBJECTS & METHODS

Patients

We evaluated all the proposed methods on a manually annotated database [21] of thousands of ten-second MER signals from 58 PD patients. The majority of them did not contain any contamination (about 75 %). The study was conducted in compliance with the Declaration of Helsinki and was approved by the Ethics Committee of the Department of Neurology, 1st Faculty of Medicine and General University Hospital in Prague. DBS electrodes were bilaterally implanted using standard stereotactic methods and incorporating intraoperative microelectrode recordings. Every participant signed an informed consent before enrolment in the study.

The sampling rate was 12 kHz and the signals were recorded with a Medtronic LeadPoint System (Medtronic, Minn., MN, USA). Internally, the signals were upsampled by a factor of 2. Digitized data was retrieved via the Leadpoint Export Utility (Medtronic, Minn., MN, USA). Band-pass filtering was applied (Butterworth 2nd order) in the 0.5–5 kHz band prior to artifact detection. The implantation procedure and domain-specific methodological details can be found in [31]. The preprocessed dataset consisted of approximately 100,000 one-second signal segments.

Our data collection process was administered by a group of trained specialists (manual). Overall, there was quite a bit of noise in the annotations (the raters did not match, there were no exact beginnings and endings of the artifacts). If an artifact emerged within 1 second, considering a fixed segmentation of the signal, the segment was marked as an artifact (+ its putative class) and each segment was ranked from one up to five raters. With more than two judges, the majority opinion was chosen (consensus). This methodology has already been described in [21].

Time-frequency representation

Feature extraction from the data could help improve the training and testing accuracies of the classifier. We considered a time-frequency representation that extracts information from the 1D MER segment (Fig.2.). Namely, using the continuous wavelet transform. A scalogram is then the absolute value of the CWT coefficients of a signal. Our requirements for the 2D representation were met by the Matlab implementation, which sets the default analytical Morse (3, 60) wavelet, where the symmetry is 3 and the time-bandwidth product is 60 [32]. The Morlet wavelet analysis has already shown appropriate properties for intracranial EEG [33], [34]. Each color JPEG image was saved to the size of 227-by-227-by-3 pixels.
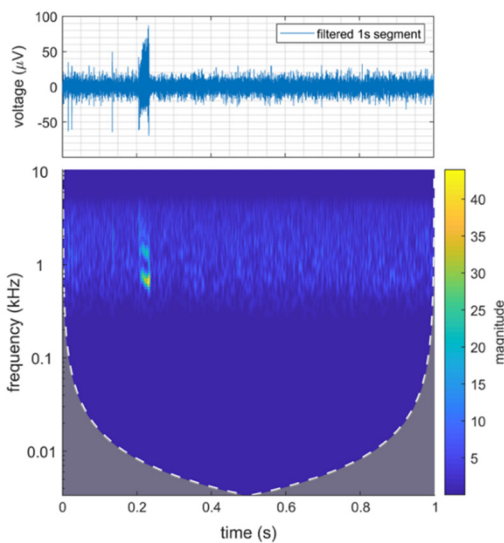

Fig.2. A one-second filtered signal segment and its corresponding 2D representation with an obvious short-living power artifact. Artifact activity is often accompanied by an increase of signal energy in particular frequency bands.

CNN architecture and settings

We employed deep transfer learning with a convolutional neural network based on AlexNet [35]. AlexNet is a CNN that is trained on more than a million images from the ImageNet database [36]. We reused the network architecture of the CNN to classify microelectrode signal segments based on images from the CWT of the time series data. Currently, the best performing neural networks were originally designed to classify images into 1000 categories. This creates major advantages for medical imaging. We were able to train the networks on smaller datasets because we were able to use pretrained models. Then, only a small weight update was needed for our specific task.

Training a deep CNN from scratch is computationally expensive and requires a large amount of training data. In various applications, a sufficient amount of training data is not available, and synthesizing new realistic training examples is not feasible. In these cases, leveraging existing neural networks that have been trained on large data sets for conceptually similar tasks is desirable. This leveraging of existing neural networks is called transfer learning. The assumption that you can train an initial model on large relevant data and transfer the hidden layers of that model to a new one, which will be developed with less focused data, is valid as we suppose that low level features are the same for all types of images. Thus, the CNN design consisted of the following steps:

• Create (taken from AlexNet) and add a new twenty-third layer including weight initialization and activation.
• Compile models including loss function (1), optimization method (2), and metrics (3), (4).
• Fit models to include learning rate, epochs, and batch size.

Essentially, the widely used optimization stochastic gradient descent algorithm might oscillate along the path of steepest descent towards the optimum. Adding a momentum term to the parameter update is a way to reduce this oscillation [37]. The stochastic gradient descent with a momentum update algorithm is employed to optimize the loss function (1):

$$\theta_{l+1} = \theta_l - \alpha\nabla L(\theta_l) + \gamma(\theta_l - \theta_{l-1}) \qquad (1)$$

where l stands for the iteration number, $\alpha > 0$ is the learning rate, θ is the parameter vector, $L(\theta_l)$ is the loss function, and γ determines the contribution of the previous gradient step to the current iteration.

The optimization algorithm finds weights and bias values that minimize the loss function (2):

$$L = -\sum_{j=1}^{C} y_j \log p_j \qquad (2)$$

where C is the number of target classes, y_j indicates that example i has label c, and p_j indicates predicted probability of class c for image x. We used the supervised method, thus we had to divide the dataset into a training, validation, and test subset, where the validation set was also used for over-fitting evaluation.

Networks were implemented in Deep Learning Toolbox environment of the Matlab 2018b. We accelerated the training processes by setting the option to run on the GPU (NVIDIA GeForce GTX 1060 6GB).

Baseline

For CNN performance benchmarking, we annotated an available in-house database with two state-of-the-art methods proposed by Bakstein et al., including default thresholds and settings [21]. Two methods, COV (autocorrelation-based approach) and MaxDiffPSD (maximum spectral difference method), were specifically reimplemented and optimized for the automatic detection of artifacts in microelectrode recordings [21], [26].

COV is a stationary segmentation method based on the variance of the signal autocorrelation function. The signal was first divided into short fixed-length segments. Subsequently, the variance ratio of the statistics of neighboring signal segments was computed. The method computes the distance matrix between all possible segment pairs and searches for the largest available component, connected by a sub-threshold path. Assuming higher stationarity in the clean signal segments, rather than in the artifact segments, the largest component is then marked as a clean signal, while the remaining signal sections are marked as artifacts. In terms of implementation, the COV classification was calculated for the entire ten-second signal and then divided into one-second segments.

MaxDiffPSD is a simpler detection method, based on the power spectral density of MER signals. The basic assumption is that the power spectral density (PSD) of a clean band-pass filtered MER signal is smooth, unlike most signals with artifacts, which commonly contain high peaks and other disturbances. In terms of implementation, MaxDiffPSD works natively with a single one-second window without the context of the entire signal.

Evaluation metrics and statistical methods

The performance of the detector was determined by a parameter that is based on the comparison of the judgment of an expert (manual) and the artifact detection results. Although expert assessment is considered as a reference gold standard, infallibility in artifact labeling cannot be expected and evaluations are strongly subjective and influenced by many factors, e.g., due to the general description of multi-trajectories of recordings, difficult differentiation of clean and neural activity with artifacts, subjective criteria of the evaluator and many more. One would expect the possibility of an expert ignoring clear artifacts or even labelling an activity that does not fully meet the criteria for the artifact. Therefore, consistent results cannot be expected, but the detectors performance should approximate the reference and also bring novelty value [38], [39].

Receiver operating characteristic (ROC) curve analysis was performed to assess diagnostic accuracy. We incorporated the R statistical software environment and pROC package [40], that is a set of tools to visualize,

smooth and compare receiver operating characteristic (ROC curves). Areas under the curve (AUC) can be compared between individual machine learning models with statistical tests based on U-statistics (DeLong's test for two correlated ROC curves) or bootstrap. Statistical tests of the hypotheses were considered significant at $\alpha = 0.05$.

The overall accuracy of the binary models was determined from the confusion matrix using the following formula (3):

$$\text{accuracy} = \frac{\text{TP+FN}}{\text{(TP+FP+TN+FN)}} \tag{3}$$

This formula represents the ratio of correctly classified segments to all the segments in the dataset (TP: True Positive, FN: False Negative, TN: True Negative, FP: False Positive).

In extending a binary metric to a multiclass problem, the data was treated as a collection of binary problems, one for each class. There are several ways to average binary metric calculations across the set of classes, each of which may be useful in some scenarios [41]. We simply calculated the macro mean of the binary metrics, giving equal weight to each class.

The proportion of correct predictions may not be useful when the two classes are of very different sizes. In addition to the accuracy and AUC-ROC, and due to the fact that the previous methods were optimized on real imbalanced data [21] we employed the Matthews Correlation Coefficient (MCC) metrics. MCC serves as a description of the optimal classifier for imbalanced data. The MCC returns value in the interval $[-1, 1]$, with 1 showing a complete agreement, -1 with a complete disagreement, and 0 showing that the prediction was uncorrelated with the ground truth [42]. MCC scores (4) were calculated via the R mccr package [43]:

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{\text{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}} \tag{4}$$

We have also further assessed the Precision-Recall (PR) curves. PR is a useful way to measure the success of the prediction when the classes are very imbalanced. These quantities are related to the F1 score, which is defined as the harmonic mean of precision and recall [41].

CNN MER validation methodology

In this study, we implemented two CNNs with the corresponding learning split-validation strategies:
• *The imbalanced case for comparison among different algorithms*. This model assumes that we classified only 2 distinct categories (clean $- 0$ vs. contaminated segment $- 1$). 5340 images were used for the training and 4930 for neural network testing, therefore 3560 images were reserved for validation and hyperparameter settings.
• *The multiclass balanced case for detection of different artifacts classes* (a finite set of species such as "power", "freq" and "baseline"). We had to redefine the last AlexNet CNN layer for 3 output classes. 567 images were used for training and 227 for neural network testing, therefore 340 images were reserved for validation and hyperparameter

settings. 378 images were available for each class representative.

In both cases, the hyperparameters finally used for offering the best results were set to 0.001 for learning rate, 0.9 for momentum, 50 training epochs and batch size of 50 samples. To finally determine what the actual classes are, the AlexNet layer has the categories of classification. Soft-labels were converted for each input as the maximum probability value.

Because of computer memory limitations, we used a subsampling of our in-house database using random selection (uniformly distributed to avoid a bias) and we covered representatives across the database for all the specific requirements for two split-validation strategies. The training phases ended after a significant decrease in validation accuracy.

3. RESULTS

In order to illustrate the performance of the deep learning framework on MER data, a series of experiments were carried out.

CNN is successful in detecting artifacts

First, we tested imbalanced benchmarking—optimized for real MERs with many clean segments. By running some benchmarking on existing models with binary predictions to see how feasible our CNN approach is (AUC-ROC = 0.75; MCC = 0.59; F1 = 0.64), we obtained a substantial performance increase (DeLong's test for two correlated ROC curves) against the results of the COV (AUC-ROC = 0.58; MCC = 0.2; F1 = 0.31) method ($p < 0.001$) and then we revealed a statistically significant improvement in comparison with MaxDiffPSD (AUC-ROC = 0.73; MCC = 0.54; F1 = 0.60; $p = 0.0497$). In all the benchmarks, CNN reported performances that were comparable and even superior to other state-of-the-art methods. The MCC metrics reflect a strong positive relationship with ground truth (+.40 to +.69) and a weak positive relationship (+.20 to +.29), respectively.
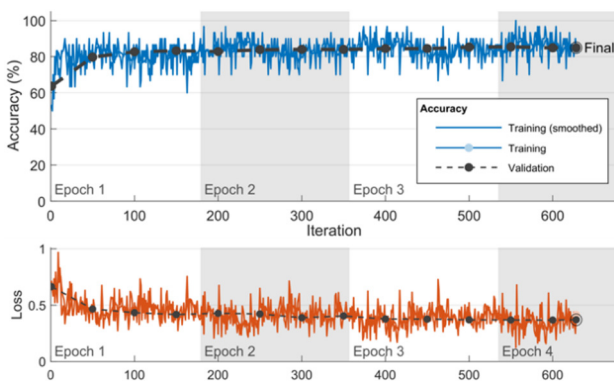
Fig.3. Preview of the learning process. Plot of model accuracy on training and validation dataset. From the plots of accuracy and loss it is evident that the model has a comparable performance on both training and validation datasets. The final total learning time of the network was very fast (in minutes). Herein it was manually stopped.

Fig.3. shows an example of the evolution of the accuracy achieved during the training phase of the AlexNet network using the validation images of our dataset.

The results from the testing datasets are shown as a confusion matrix for all the artifact detectors (Fig.4.). The CNN based approach was shown to have an accuracy of 88.1 %.
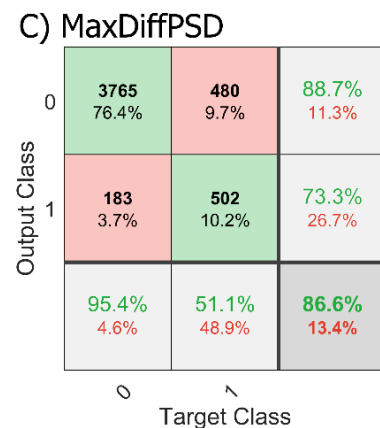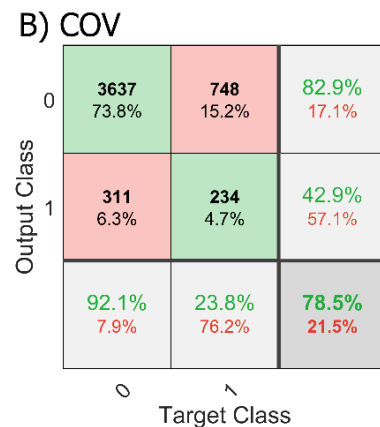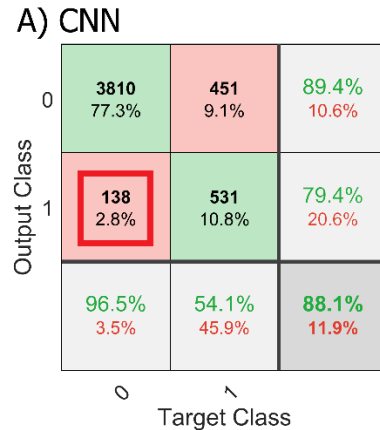
Fig.4. Confusion matrices for imbalanced cases show the distribution of the test segments into false/true positives/negatives. A summary of the statistics is given in the last rows/columns. A: CNN. The neural network can identify candidates for artifacts which could have been missed (red box). B: COV. C: MaxDiffPSD.

Figures of ROC and PR curves and their areas including probabilistic CNN soft-labels are presented in Fig.5. and Fig.6.
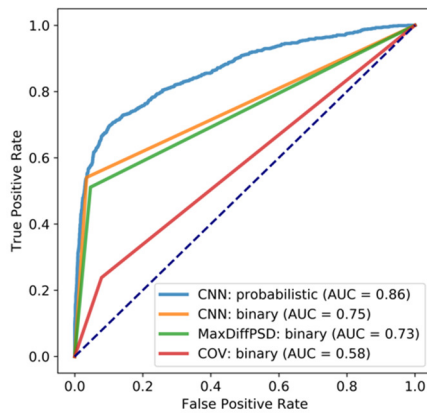

Fig.5. ROC curves for individual models to evaluate output quality. Binary inputs give one point on the ROC curve while the probabilistic CNN returns a smooth curve where an optimal point could be found.
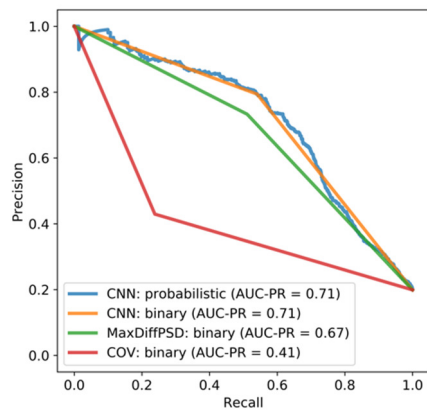

Fig.6. Precision-Recall curves. Unlike the ROC curve, PR curves are very sensitive to imbalance. A classifier optimized for good AUC-ROC on an unbalanced data is likely to obtain poor PR results.

Although a high computational cost is typical for deep learning architectures during training, the evaluation phase is very quick and efficient. Processing of 1-s MER segment took on average 0.885 (std. 0.039) seconds for the CWT computation and 0.005 (std. 0.0004) seconds for the CNN evaluation on a standard i5 2.8 GHz PC with 8 GB RAM (Table 1.).

Table 1. Descriptive statistics that summarize the speed of computation in seconds (4930 images).

Statistic	CWT	CNN	CWT+CNN
mean	0.8850	0.0048	0.8897
standard deviation	0.0385	0.0004	0.0385
min	0.7978	0.0045	0.8025
50% percentile	0.8863	0.0047	0.8910
max	2.3352	0.0126	2.3399

CNN can distinguish among classes of artifacts

As a second result, we determined balanced-data multiclass CNN results with a prediction accuracy of 75.3 %, which is a noticeably good level and clearly reveals the effectiveness of the proposed method as presented in Fig.7. It was observed that both power and baseline groups had comparable accuracy values. The biggest difference was given for the frequency label. Frequency and baseline artifacts were incorrectly labelled as a power contamination several times. By accessing "soft-labels", the user can adjust the sensitivity for each type of artifact, so that there is a higher sensitivity for the types that are more troublesome.

Fig.7. Multiclass classification using a CNN detector on the testing set.

Detailed focus on artifacts

Each layer of a CNN produces a response, or activation, to an input image. However, there are only a few layers within the CNN that are suitable for image feature extraction. The layers at the beginning of the network capture basic image features, such as edges and blobs. We examined the activations and discovered which features AlexNet learned by comparing areas of activation with the original image. Our method detected artifacts and focused on the most activated areas (Fig.8.).

Table 2. Example measurements for the set of properties for each recognized artifact object in the binary image.

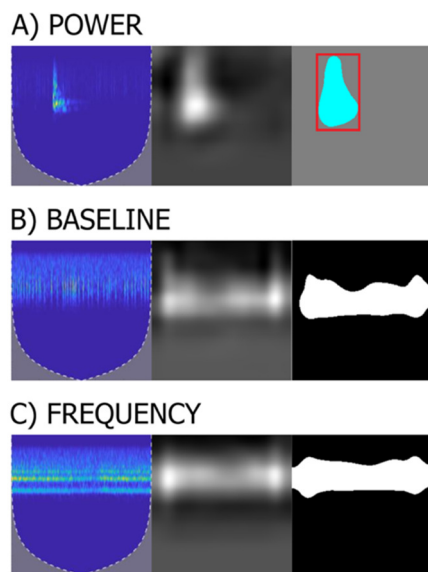| Feature	Value (mostly pixel-based)
Centroid	[71.94 83.45]
BoundingBox	[43.50 16.50 64 115]
MajorAxisLength	115.59
MinorAxisLength	57.14
Area	4622
Eccentricity	0.87
Orientation	-85.03
ConvexArea	5055
Perimeter	285.60

Fig.8. Our methodology was able to find the strongest channel for a particular MER segment and mark the exact artifact onset time. Herein, we compared the strongest channel with the original image. Information retrieval was performed on the strongest channel in the 5th Convolutional Layer. Using the automatic Otsu's global threshold segmentation, image parameters are extracted (red box) from the artifact (Table 2.).

4. DISCUSSION

This paper applies deep learning strategy to classifying artifacts in Parkinson's disease microelectrode recordings. This method could be of potential interest not only for invasive recordings but also for noninvasive ones (e.g., electroencephalography). The data was first preprocessed via wavelet transformation and then analyzed using convolutional neural network.

Evaluation criteria showed that the proposed algorithms obtained the following knowledge:

1) CNN results are at least comparable to the previous simple methods (Fig.4., Fig.5., and Fig.6.).

2) Individual classes of artifacts are recognizable (Fig.7.).

3) Our method implemented the active layer highlighting as feedback so that an expert can make better decision (Fig.8.).

Our novel approach is benchmarked on a gold-standard dataset and is supported by experimental validation. The performance of the CNN algorithm was compared with existing methods (MaxDiffPSD, COV) and relevant statistics are given (e.g., AUC-ROC, AUC-PR, accuracy, MCC). The results indicated that all metrics were highest for CNN. Indeed, the simplest method MaxDiffPSD achieved almost the same performance as the deep learning (CNN) method. But, in addition to the ability to successfully classify, our network can also decide which class of artifact is present (multiclass problem). This is a respectable result if one considers that the task was quite complex. In this regard, only artifact classes of interest can be selected/classified, which is a real new feature and advantage over the state-of-the art methods and the article [21].

Subsequently, we confirmed our hypothesis of its capacity to report the most activated layer and to localize the area of the MER in which artifacts originate. The authors of the AlexNet stated in their original paper that their model is capable of identifying the region of interest on the image [44]. Therefore, we have adapted this proven model for our problem. Our approach uniquely allows for targeting the time location of the artifact in a one-second segment, which has not been previously designed and implemented. To accurately evaluate the onset/offset of each identified artifact, we do not have labeled data for objective validation, nor a means to aggregate the 2D binary matrix into one dimension. So, this feature cannot be compared with identification based on state-of-the-art methods (e.g., MaxDiffPSD). Moreover, this is even beyond the capacity of state-of-the-art methods.

The one type of existing state-of-the-art methods for artifact detection in single-channel MER that has been rigorously tested is unsupervised change-point detection [23]-[25]. The authors presented the extremely high detection accuracies of change-point detection, which were almost 100 % on simulated data. We show that real-world validation performance may not be as robust when the problem is evaluated as supervised (COV accuracy = 78.5 %). Among others, the COV method typically labels a segment as an artifact when an artifact was not present (a common case for a large sudden spike). Such problems can be overcome by a CNN.

Offline analysis is of no use to neurosurgeons. A quick (within 0.5-1.0 second) result of the calculation is desirable during the MER recording in the surgery room. Our solution meets these requirements with processing speeds below 0.9 seconds (Table 1.). A time analysis was performed on all test data (4930 images) and classification by the saved model for the new image was very fast. As shown, it takes nearly 1 second to calculate the CWT and resize the image. The proposed techniques could be implemented during surgery for near real-time applications of detecting artifact classes or potential biomarkers. CWT + CNN seems to be an effective way to extract hidden information from a MER in DBS.

Research in the field of DBS-MER processing has historically been limited due to low volumes of annotated data. Despite this fact, we have a large and comprehensive dataset, gathered from electrodes implanted in the brain, offering a lot of randomly selected trajectories of ten-second long brain activity recordings. In the presented problem, we have thousands of training samples and there were no obvious effects of overfitting. This method appears to be transferable to other datasets and we plan to apply it in dystonia [31].

This paper presents several evident advances but also contains limitations. We have evaluated several combinations of the hyperparameters of the CNN to be tuned, but we have not systematically performed the hyperparameter optimization. In this regard, we have only retrained AlexNet for 2 resp. 3 output classes in the last layer. This has allowed for a small subset of weights to be changed, while the majority of the initially optimized core

AlexNet weights were retained, so we leveraged the AlexNet optimization process. We would rather refer to how the hyperparameters were systematically optimized in the original architecture [44]. Nevertheless, a deeper analysis of the hyperparameters can lead to even better results [45]. It is apparent that the loss functions decreased during training and we captured the artifact patterns successfully.

AlexNet is already dated architecture. For further research, we would consider novel deep learning architectures such as Inception [46], residual networks [47], or dense-net [48] to improve the results. It has been shown that other architecture and resizing of the input image may have a significant effect on the accuracy of the classification and the current satisfactory accuracy could increase by more than 10 % [49].

In the future, we offer a semi-supervised principle, namely submitting vague images for re-evaluation by the rater. We identified the active learning approach as a viable route for further research [50].

5. CONCLUSION

Currently, there is a variety of software methods that can enable detection of DBS biomarkers. Such software will enable spike sorting, spike clustering, and network synchronization analysis. Therefore, the concept of microrecording seems very attractive and promising in the tailoring of individual patient therapy. This article illustrates an innovative approach to microelectrode recordings processing. The main result is the implementation of an artifact detector based on the convolutional neural network. To our knowledge, deep learning in the field of DBS has not been studied yet. This automatic approach can lead to improved DBS diagnostic procedures and can improve DBS effectiveness. The results suggest that the presented methodology is promising for DBS signal processing evaluation and is able to capture more hidden patterns in MER data, specifically not only artifacts, but also, for example, some other characteristic features (neuronal firing or even identifying potential candidates for biomarkers). It can be applied to any movement disorders in deep brain stimulation procedure implantation. Models, source codes and anonymized mock data are available upon request.

REFERENCES

[1] Bouton, C. (2017). Cracking the neural code, treating paralysis and the future of bioelectronic medicine. *Journal of Internal Medicine*, 282 (1), 37-45.

[2] Wichmann, T., DeLong, M.R. (2016). Deep brain stimulation for movement disorders of basal ganglia origin: Restoring function or functionality? *Neurotherapeutics*, 13 (2), 264-283.

[3] Kocabicak, E., Alptekin, O., Ackermans, L., Kubben, P., Kuijf, M., Kurt, E., et al. (2015). Is there still need for microelectrode recording now the subthalamic nucleus can be well visualized with high field and ultrahigh MR imaging? *Frontiers in Integrative Neuroscience*, 9, 46.

[4] Abosch, A., Timmermann, L., Bartley, S., Rietkerk, H. G., Whiting, D., Connolly, P.J., et al. (2013). An international survey of deep brain stimulation procedural steps. *Stereotactic and Functional Neurosurgery*, 91 (1), 1-11.

[5] Bour, L.J., Contarino, M.F., Foncke, E.M.J., de Bie, R.M.A., van den Munckhof, P., Speelman, J.D., Schuurman, P.R. (2010). Long-term experience with intraoperative microrecording during DBS neurosurgery in STN and GPi. *Acta Neurochirurgica*, 152 (12), 2069–2077.

[6] Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., et al. (2019). A guide to deep learning in healthcare. *Nature Medicine*, 25 (1), 24–29.

[7] Al-Ajlan, A., El Allali, A. (2018). CNN-MGP: Convolutional neural networks for metagenomics gene prediction. *Interdisciplinary Sciences: Computational Life Sciences*, doi.org/10.1007/s12539-018-0313-4.

[8] Bursa, M., Lhotska, L. (2017). The use of convolutional neural networks in biomedical data processing. In *Information Technology in Bio- and Medical Informatics*. Springer, 100-119.

[9] Kietzmann, T.C., McClure, P., Kriegeskorte, N. (2018). Deep neural networks in computational neuroscience. *bioRxiv*, doi.org/10.1101/133504.

[10] Vásquez-Correa, J.C., Orozco-Arroyave, J.R., Nöth, E. (2017). Convolutional neural network to model articulation impairments in patients with Parkinson's disease. In *Interspeech 2017*. ISCA, 314-318, DOI: 10.21437/Interspeech.2017-1078.

[11] Cote-Allard, U., Fall, C.L., Drouin, A., Campeau-Lecours, A., Gosselin, C., Glette, K., et al. (2019). Deep learning for electromyographic hand gesture signal classification using transfer learning. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 27 (4), 760-771.

[12] Acharya, U.R., Oh, S.L., Hagiwara, Y., Tan, J.H., Adeli, H. (2018). Deep convolutional neural network for the automated detection and diagnosis of seizure using EEG signals. *Computers in Biology and Medicine*, 100, 270-278.

[13] Al Rahhal, M.M., Bazi, Y., Al Zuair, M., Othman, E., BenJdira, B. (2018). Convolutional neural networks for electrocardiogram classification. *Journal of Medical and Biological Engineering*, 38 (6), 1014–1025.

[14] Yıldırım, Ö., Pławiak, P., Tan, R.-S., Acharya, U.R. (2018). Arrhythmia detection using deep convolutional neural network with long duration ECG signals. *Computers in Biology and Medicine*, 102, 411–420.

[15] Yildirim, Ö. (2018). A novel wavelet sequence based on deep bidirectional LSTM network model for ECG signal classification. *Computers in Biology and Medicine*, 96, 189–202.

[16] The MathWorks, Inc. *Deep Learning Toolbox: Create, analyze, and train deep learning networks.* www.mathworks.com/products/deep-learning.html.

[17] *TensorFlow: An end-to-end open source machine learning platform.* https://www.tensorflow.org/.

[18] *Keras Documentation: The Python deep learning library.* https://keras.io/.

[19] *PyTorch: An open source machine learning framework that accelerates the path from research prototyping to production deployment.* https://www.pytorch.org.

[20] *Caffe2: A new lightweight, modular, and scalable deep learning framework.* http://caffe2.ai/.

[21] Bakštein, E., Sieger, T., Wild, J., Novák, D., Schneider, J., Vostatek, P., et al. (2017). Methods for automatic detection of artifacts in microelectrode recordings. *Journal of Neuroscience Methods*, 290, 39–51.

[22] Cagnan, H., Dolan, K., He, X., Contarino, M.F., Schuurman, R., van den Munckhof, P., et al. (2011). Automatic subthalamic nucleus detection from microelectrode recordings based on noise level and neuronal activity. *Journal of Neural Engineering*, 8 (4), 046006.

[23] Falkenberg, J.H., McNames, J., Aboy, M., Burchiel, K.J. (2003). Segmentation of extracellular microelectrode recordings with equal power. In *Proceedings of the 25th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2475-2478.

[24] Aboy, M., Falkenberg, J.H. (2006). An automatic algorithm for stationary segmentation of extracellular microelectrode recordings. *Medical & Biological Engineering & Computing*, 44 (6), 511-515.

[25] Guarnizo, C., Orozco, A., Castellanos, G. (2008). Microelectrode signals segmentation using stationary wavelet transform. In *2008 International Conference on BioMedical Engineering and Informatics*. IEEE, 450-454.

[26] Bakstein, E., Schneider, J., Sieger, T., Novak, D., Wild, J., Jech, R. (2015). Supervised segmentation of microelectrode recording artifacts using power spectral density. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 1524-1527.

[27] Weegink, K.J., Bellette, P.A., Varghese, J.J., Silburn, P.A., Meehan, P.A., Bradley, A.P. (2017). A parametric simulation of neuronal noise from microelectrode recordings. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25 (1), 4-13.

[28] Moran, A., Bar-Gad, I., Bergman, H., Israel, Z. (2006). Real-time refinement of subthalamic nucleus targeting using Bayesian decision-making on the root mean square measure. *Movement Disorders*, 21 (9), 1425-1431.

[29] Dolan, K., Martens, H.C.F., Schuurman, P.R., Bour, L.J. (2009). Automatic noise-level detection for extra-cellular micro-electrode recordings. *Medical & Biological Engineering & Computing*, 47 (7), 791-800.

[30] The MathWorks, Inc. *Convolutional Neural Network.* https://www.mathworks.com/solutions/deep-learning/convolutional-neural-network.html.

[31] Klempir, O., Krupicka, R., Cejka, V., Jech, R. (2018). Microelectrode neuronal activity biomarker of the internal globus pallidus in dystonia correlates with long-term neuromodulation effects. In *2018 IEEE 20th International Conference on e-Health Networking, Applications and Services (Healthcom)*. IEEE, 1-6.

[32] Lilly, J.M. (2017). Element analysis: A wavelet-based method for analysing time-localized events in noisy time series. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Science*, 473 (2200), 20160776.

[33] Gadhoumi, K., Lina, J.-M., Gotman, J. (2012). Discriminating preictal and interictal states in patients with temporal lobe epilepsy using wavelet analysis of intracerebral EEG. *Clinical Neurophysiology*, 123 (10), 1906-1916.

[34] Gadhoumi, K., Lina, J.-M., Gotman, J. (2013). Seizure prediction in patients with mesial temporal lobe epilepsy using EEG measures of state similarity. *Clinical Neurophysiology*, 124 (9), 1745-1754.

[35] Sharma, N., Jain, V., Mishra, A. (2018). An analysis of convolutional neural networks for image classification. *Procedia Computer Science*, 132, 377–384.

[36] *ImageNet.* http://www.image-net.org.

[37] Murphy, K.P. (2012). *Machine Learning: A Probabilistic Perspective*. MIT Press.

[38] Jiruska, P., Finnerty, G.T., Powell, A.D., Lofti, N., Cmejla, R., Jefferys, J.G.R. (2010). Epileptic high-frequency network activity in a model of non-lesional temporal lobe epilepsy. *Brain*, 133 (5), 1380–1390.

[39] Casson, A.J., Luna, E., Rodriguez-Villegas, E. (2009). Performance metrics for the accurate characterisation of interictal spike detection algorithms. *Journal of Neuroscience Methods*, 177 (2), 479–487.

[40] Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., Müller, M. (2011). pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, 12, 77.

[41] *Model Evaluation: Quantifying the quality of predictions.* https://scikit-learn.org/stable/modules/model_evaluation.html.

[42] Boughorbel, S., Jarray, F., El-Anbari, M. (2017). Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. *PLoS One*, 12 (6), e0177678.

[43] Iuchi, H. (2017). *mccr: The Matthews correlation coefficient (v. 0.4.4).* https://CRAN.R-project.org/package=mccr.

[44] Krizhevsky, A., Sutskever, I., Hinton, G.E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60 (6), 84-90.

[45] Montavon, G., Orr, G.B., Müller, K.-R. (eds.) (2012). *Neural Networks: Tricks of the Trade*. Springer.

[46] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., et al. (2015). Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 1-9.

[47] Talo, M., Baloglu, U.B., Yıldırım, Ö., Rajendra Acharya, U. (2019). Application of deep transfer learning for automated brain abnormality classification using MR images. *Cognitive Systems Research*, 54, 176-188.

[48] Arsalan, M., Naqvi, R., Kim, D., Nguyen, P., Owais, M., Park, K. (2018). IrisDenseNet: Robust iris segmentation using densely connected fully convolutional networks in the images by visible light and near-infrared light camera sensors. *Sensors*, 18 (5).

[49] Llamas, J.M. Lerones, P.M., Medina, R., Zalama, E., Gómez-García-Bermejo, J. (2017). Classification of architectural heritage images using deep learning techniques. *Applied Sciences*, 7 (10).

[50] *How to retrain an image classifier for new categories*. www.tensorflow.org/hub/tutorials/image_retraining.