Uncertainty-Aware Bayesian Deep Learning with Noisy Training Labels for Epileptic Seizure Detection

Deeksha M. Shama[1,2] and Archana Venkataraman[1,2]

[1] Dept. of Electrical and Computer Eng., Johns Hopkins University, Baltimore, USA
[2] Dept. of Electrical and Computer Eng., Boston University, Boston, USA
dshama1@jhu.edu, archanav@bu.edu

Abstract. Supervised learning has become the dominant paradigm in computer-aided diagnosis. Generally, these methods assume that the training labels represent "ground truth" information about the target phenomena. In actuality, the labels, often derived from human annotations, are noisy/unreliable. This *aleoteric uncertainty* poses significant challenges for modalities such as electroencephalography (EEG), in which "ground truth" is difficult to ascertain without invasive experiments. In this paper, we propose a novel Bayesian framework to mitigate the effects of aleoteric label uncertainty in the context of supervised deep learning. Our target application is EEG-based epileptic seizure detection. Our framework, called BUNDL, leverages domain knowledge to design a posterior distribution for the (unknown) "clean labels" that automatically adjusts based on the data uncertainty. Crucially, BUNDL can be wrapped around any existing detection model and trained using a novel KL divergence-based loss function. We validate BUNDL on both a simulated EEG dataset and the Temple University Hospital (TUH) corpus using three state-of-the-art deep networks. In all cases, BUNDL improves seizure detection performance over existing noise mitigation strategies.

Keywords: aleoteric Label Noise · Trustworthy AI · EEG · Epilepsy

1 Introduction

Deep learning is becoming ubiquitous in computer-aided diagnosis with medical imaging data. The most common paradigm is *supervised learning*, which relies on labeled training data to learn complex and generalizable patterns. The underlying assumption of supervised learning is that the training labels provide accurate"ground truth" information about the phenomena of interest. In actuality, these labels can be unreliable due to noise, human error, and subjective biases. This unreliability can be viewed as *aleoteric uncertainty* arising from the complex imaging data and the manual label annotations. Aleoteric uncertainty is further exacerbated in modalities that are not easily readable through visual inspection, such as genomics data, fMRI connectomics, and electroencephalography (EEG). Particularly, epileptic seizure detection using EEG has

seen extensive adoption of supervised AI models [27,2,5,4,19,9,14] but also has high aleoteric uncertainty due to environmental noise and data artifacts [6,3]. As a consequence, the manually annotated seizure interval and onset location are ambiguous, and AI models trained on such noisy labels inadvertently learn misleading features, resulting in poor generalization performance.

Considerable progress has been made within the AI community to quantify uncertainty [1,12,13] and to adaptively de-noise the data while still trusting the given labels [20]. More recently, there has been work that tackles the problem of *learning with uncertain labels* [23]. These methods can be broadly grouped into three categories. The first category is data pruning, in which a pretraining step is used to re-weight or remove noisy samples prior to training [26,16,17]. These methods can lead to degeneracies in complex data modalities by removing critical data points. The second category proposes "robust architectures" that use an auxiliary deep network to estimate the transition from clean label to noisy label via a joint optimization [7], EM strategies [15], or sequentially trained multi-networks [25]. These methods suffer from increased time complexity, scalability issues, and over-fitting on smaller datasets. The third category devises "robust loss functions" which use prior information for self-supervised label adjustment [10], model regularization to prevent overfitting [28], and loss correction via bootstrapping [18,21]. However, these methods do not model instance-dependent label transitions, which is often the case for complex data modalities.

We propose a novel statistical framework called Bayesian UNcertainty-aware Deep Learning (BUNDL) to mitigate the impact of label inaccuracies when training deep networks. BUNDL assumes a conditional Bernoulli distribution for the unknown "clean labels" that alters the predicted posterior probability when uncertainty is higher in the data. Our target application is epileptic seizure detection from scalp EEG data. Accordingly, we leverage domain knowledge that the seizure is generally over-segmented by clinicians [3,8] and infer the clean label posterior from the observed EEG by minimizing a novel KL divergence loss function. BUNDL is developed in a network-agnostic manner and is validated on both simulated EEG data and the publicly available Temple University Hospital (TUH) corpus [22] using *three state-of-the-art deep networks* for seizure detection. Our results show that BUNDL achieves better seizure detection performance than existing noise mitigation strategies while being network-independent and easily adapted to new applications. We also conduct a secondary analysis of seizure onset zone (SOZ) localization and demonstrate that using BUNDL to "correct" the seizure detection task also improves localization.

2 Unified Framework to Mitigate Aleoteric Label Noise

Fig. 1(left) illustrates the graphical model that underlies BUNDL. For simplicity, we assume label corruption depends only on the current input, not the entire training dataset. Mathematically, we let $x \in \mathbb{R}^{N_d}$ be a random variable for the input data of dimension N_d, $\hat{y} \in \{0, 1\}$ be a Bernoulli random variable corresponding to the given noisy label, and $y \in \{0, 1\}$ be the unobserved clean

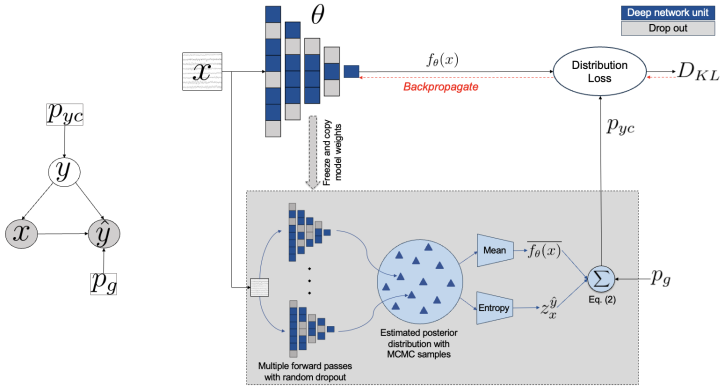

Fig. 1: **Left:** Directed graphical model during training. **Right:** Overall training strategy of BUNDL using the observed data x and given labels $\hat{y}$ in every epoch.

label. In our case, x is a single frame of multichannel EEG, $\hat{y}$ is the corresponding clinician annotated seizure versus baseline classification, and y is the unobserved ground truth seizure information. Our generative process(Fig. 1(left)) assumes that the clean labels y are used to generate the input data x and that both x and y inform the noisy labels $\hat{y}$. Finally, both x and $\hat{y}$ are observed during training.

2.1 Learning the Clean Label Posterior via Distributional Loss

Our goal during modified supervised training is to learn a function $f_\theta(\cdot)$ parameterized by θ that infers the clean label posterior distribution given the input data, i.e., $p(y|x)$. This function can then be applied to new test data samples.

We train $f_\theta(\cdot)$ using the labeled pairs $(x, \hat{y})$ by first defining the noisy label posterior distribution as $p(\hat{y}|x) = \mathcal{B}(\hat{y}; p_g)$, where $\mathcal{B}(\cdot)$ denotes a Bernoulli distribution, and the parameter p_g is the clinician provided (binary) label of baseline versus seizure activity. We note that if the labels were instead provided as a continuous measure of clinician confidence, then these could be used as p_g.

We account for label uncertainty by defining the posterior distribution of the clean label y given the observed data and noisy label pair $(x, \hat{y})$ as follows:

$$p(y|x, \hat{y}) = \mathcal{B}(y; p_g(1 - z_x^{\hat{y}}) + \overline{f_\theta(x)} z_x^{\hat{y}}), \tag{1}$$

where $z_x^{\hat{y}}$ is instance-dependent aleoteric uncertainty and $\overline{f_\theta(x)}$ is the average of the model prediction $f_\theta(\cdot)$ across multiple stochastic forward passes to the model using dropout. Intuitively, when $z_x^{\hat{y}}$ is low, the model trusts the given labels p_g since accuracy of manual annotations of clean samples is reliable. When $z_x^{\hat{y}}$ is high, the model hinges towards its self-supervision $\overline{f_\theta(x)}$.

By marginalizing $\hat{y}$ in Eq. (1), we obtain the clean label posterior $p(y|x) = \mathcal{B}(y; p_{yc})$, where the Bernoulli parameter p_{yc} is computed as follows:

$$p_{yc} = p_g(z_x^1 \cdot \overline{f_\theta(x)} + p_g(1 - z_x^1)) + (1 - p_g)(z_x^0 \cdot \overline{f_\theta(x)} + p_g(1 - z_x^0)) \tag{2}$$

We implement the function $f_\theta(\cdot)$ using a deep neural network. The optimal parameter θ^* are learned by minimizing the KL divergence between the Bernoulli distribution parameterized by Eq. (2) and the deep network prediction $Q(y|x;\theta) = \mathcal{B}(y; f_\theta(x))$ for all samples in the training dataset:

$$\begin{aligned}\theta^* &= \arg\min_\theta D_{KL}(P||Q) \\ &= \arg\max_\theta \Big(p_{yc} \cdot log(f_\theta(x)) + (1 - p_{yc}) \cdot log(1 - f_\theta(x))\Big)\end{aligned}$$

Computing Uncertainty from Monte-Carlo Dropout (MCD): As shown in Fig. 1(right), BUNDL estimates the posterior distribution of the clean labels through multiple forward passes of the input EEG to the deep network. These passes generate "posterior estimates" with the randomness provided by dropout [24]. We estimate the data uncertainty $z_x^{\hat{y}}$ as the empirical entropy (rescaled between 0 and 1) of these posterior samples [12]. From a practical standpoint, the network output is random in the initial stages of training and cannot provide a good uncertainty estimate. Therefore, we pretrain the network with a subset of data away from the seizure onset and set $z_x^{\hat{y}} = 0$. This way we avoid learning from highly uncertain EEG data around change points in early stages of training. We then update the network parameters using the $z_x^{\hat{y}}$ continuously predicted by the model for another 30 epochs. We also use a small tolerance of 0.001 for p_g to ensure numerical stability. The dropout rate is set at 0.2 and 20 posterior samples are used in every epoch. The learning rate is chosen through grid search in the range $[10^{-6}, 10^{-2}]$ to choose the best performing rate across nested 10-fold cross validation.See supplement for the algorithms.

Domain Knowledge for Seizure Detection: The seizure interval tends to be "over-segmented" by clinicians due to the comparatively higher cost of missing a seizure than introducing false positives [3,8]. Consequently, we opt to trust the control class labels and set $z_x^0 \doteq 0$ during training only. The overestimated seizure class is unreliable, hence, we quantify z_x^1 using the entropy of posterior samples. Hence, $p_{yc} = p_g(z_x^1 \overline{f_\theta(x)} + p_g(1 - z_x^1))$ [3].

2.2 Evaluation setup

We use the optimal parameters θ^* learned for $f_\theta(\cdot)$ during training to directly estimate the clean label parameter p_{yc} from new testing data. We report metrics at the window and seizure levels for detailed comparison. At window level, we compare area under the receiver operating characteristic curve (AUROC), and area under the precision-recall curve (AUPRC) aggregated across individual time points. During cross validation, we select the learning rate and detection threshold for each method to allow no more than 3 minutes of false positives per hour in the validation dataset. At the seizure level, we report the false positive rate (FPR) (min/hour), sensitivity, and the latency (seconds) of seizure prediction. We perform repeated 10-fold cross validation with patient-level splits

[3] $(1 - p_g)(p_g(1 - z_x^0) + \overline{f_\theta(x)} z_x^0) \doteq 0$ when $p_g \in \{0, 1\}$

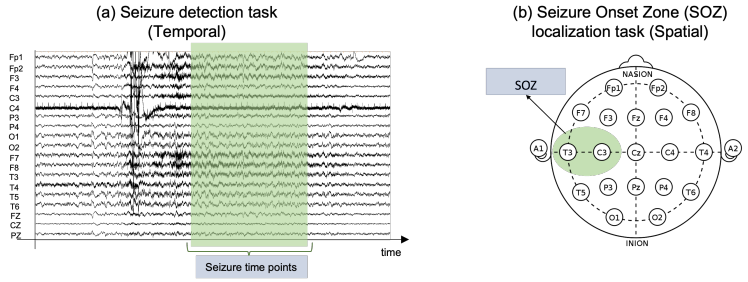

Fig. 2: (a) Example EEG from 19 channel 10-20 montage plotted across time. Seizure interval is denoted in green. (b) Corresponding scalp plot depicting SOZ

Baseline Algorithms: We compare BUNDL with two state-of-the-art approaches to handle noisy labels: (1) Self Adaptive learning (SelfAdapt) [10], which alters the clean label posterior using self-supervision independent of data uncertainty, and (2) Noisy Adaptation Layer (NAL) [7], which builds a robust model to jointly estimate clean label and noisy label posterior using EM like algorithm. We use their "complex model" as it is better suited to the task. Finally, our baseline is the traditional cross-entropy loss (CEL), which does not assume any label noise. We use the state-of-art DeepSOZ model [14] for all comparisons.

Network Comparison: Since BUNDL is model-agnostic, we use three recently published networks for seizure detection to compare the performance of BUNDL against traditional CEL-based training. These networks are as following:

1. DeeSOZ [14] consists of a transformer that employs self-attention to extract features across EEG channels in a time window. The global features from the transformer is then processed by an LSTM to predict seizure, whereas channel-wise features are passed through pooling layer to localize SOZ.
2. TGCN [4] employs eight layers of spatio-temporal convolutions (STC) consisting of 1-D convolutions to process EEG within each channel's neighborhood. This is followed by three linear layers to predict seizure
3. CNN [5] also consists of 1-D convolutions and batch normalization that process an EEG time window altogether unlike TGCN that does it within a neighborhood. This is followed by LSTM layer to predict seizure activity.

Extension to Seizure Onset Zone (SOZ) Localization:. SOZ localization is crucial for real-world epilepsy management. Unlike seizure detection, the goal of SOZ localization is to identify the EEG channels from which the seizure activity originates, as shown in Fig.2. We hypothesize that accounting for noise in the seizure detection labels will help us stabilize and/or improve SOZ localization. We compare the performance of DeepSOZ for SOZ localization when the seizure detection branch is trained with BUNDL to when it is trained with CEL. We report accuracy averaged at seizure-level and patient-level as in [14].

Table 1: Description of patient demographics in simulated and TUH dataset.

	Simulated dataset	TUH dataset
Number of patients	120	120
Total Number of seizures	631	642
Average seizures per patient	5.3±2.69	14.7±25.2
Min/Max seizures per patient	1/10	1/152
Average EEG duration per patient	52.8±26.97 min	79.8±135 min
Average seizure duration	170.5±120.4 sec	88.0±123.5 sec
Min/Max seizure duration	29/491 sec	7.5/1121 sec

3 Results

3.1 Epilepsy Datasets

Simulated Data:[4] We use the SEREEGA framework [11] in MATLAB to simulate 10 min EEG recordings for 120 subjects with 1-10 seizure events per subject. We select 10-20 montage channels from the 32-channel New York Head (ICBM-NY) leadfield matrix. Background activity is generated via four sources in each brain quadrant with continuous ERSP data in alpha, beta, and occasional slow waves, and one randomly located source for spike artifacts (6-14 Hz). Background sources include multiple randomly chosen white noise intervals. Seizure source locations vary for each subject, with continuous intervals of spikes and sharp waves (2.5-4 Hz) having uniformly varying onset times and duration. We generate noisy labels by oversegmenting seizures by 30% for training and using the clean labels for evaluation. We choose to oversegment seizures, as this matches our clinical hypotheses, and we defer other types of label noise to future.

TUH Corpus:[5] TUH [22] is a public dataset with 642 seizure recordings from 120 patients (Table 1. Each recording contains 10-20 montage EEG data and clinician-annotated intervals of the temporal seizure interval and the seizure onset channels. For preprocessing, we resample the EEG signals to 200 Hz to ensure uniformity. We filter the signals (1.6-30 Hz) and apply a clipping mechanism to remove high-intensity artifacts, setting the threshold at two standard deviations from the mean. All signals are normalized to have zero mean and unit variance within a recording. We crop recordings to a duration of 10 minutes centered around the seizure interval, while ensuring a *uniform distribution of onset times* and segment the input EEG to one-second non-overlapping windows.

3.2 Performance Comparisons with BUNDL

Seizure Detection Accuracy: Table 2 compares BUNDL using DeepSOZ with three state-of-the-art noise mitigation strategies on the simulated data. The performance metrics are computed against the ground-truth *clean labels* for simulated data. We observe that BUNDL achieves the best overall performance with

[4] Simulated data and code repository
[5] TUH dataset.

Table 2: Seizure detection performance based on the DeepSOZ architecture [14] under different noise mitigation strategies on **simulated data**.

Method	Window-Level		Seizure-Level		
	AUROC	AUPRC	FPR	Sensitivity	Latency
BUNDL	**0.97±0.02**	**0.95±0.03**	**2.60±0.85**	**0.90±0.06**	**30.1±21.7**
SelfAdapt [10]	0.93±0.03	0.89±0.04	5.63±2.17	0.84±0.05	59.9±31.7
NAL [7]	0.95±0.03	0.92±0.03	4.88±1.61	0.89± 0.07	56.0± 22.4
CEL [14]	0.93±0.04	0.89±0.04	5.8±3.6	0.85±0.10	58.4±34.3

Table 3: Seizure detection performance based on the DeepSOZ architecture [14] under different noise mitigation strategies on **TUH data**.

Method	Window-Level		Seizure-Level		
	AUROC	AUPRC	FPR	Sensitivity	Latency
BUNDL	**0.917±0.031**	**0.60±0.15**	**2.30±1.10**	0.742±0.08	**7.6±10.9**
SelfAdapt [10]	0.908±0.030	0.55±0.14	2.50±1.23	0.768±0.069	17.7±14.8
NAL [7]	0.912±0.034	0.60±0.15	4.24±1.96	**0.829± 0.056**	22.3± 15.1
CEL [14]	0.916±0.028	0.60±0.14	3.6±1.77	0.786±0.065	16.9±16.2

a mean AUROC of 0.97 and a mean AUPRC of 0.95 against the clean labels at the one-second window level. While NAL achieves similar window-level performance, BUNDL shows considerable gains at the seizure level. Specifically, BUNDL reduces both latency and false-positive rate (FPR) by 50% over NAL. The SelfAdapt and CEL methods have similar overall performance, which indicates that SelfAdapt is failing to capture and/or mitigate label noise.

Table 3 reports the seizure detection accuracy of all four methods (BUNDL and baselines) on the TUH dataset. In this case, the performance metrics are computed against the (presumed noisy) *given labels*, as we do not have access to "clean labels" for real-world data. This mismatch between the model outputs (i.e., estimated clean labels) and the given labels results in a uniform decrease in window-level AUROC and AUPRC as compared to the simulated data. At the seizure level, BUNDL has the lowest FPR (2.3 min/hour) and onset latency (7.6 sec), which is a significant improvement over the baseline methods. We note that the baseline methods achieve a slightly higher sensitivity than BUNDL. This can be partially attributed to the mismatched evaluation against the noisy given labels. In addition, the increased FPR for the baselines suggest that they are over-segmenting the seizure, which can also lead to higher sensitivity.

Fig. 3 illustrates the predicted "clean" seizure probability p_{yc} across time for EEG from a patient in TUH. BUNDL correctly predicts the seizure whereas other methods have increased false positives (red) and latency.

Applying BUNDL to Different DL Models: Table 4 compares BUNDL with CEL (i.e., no noise mitigation) using three state-of-the-art deep networks. In simulated data, BUNDL is uniformly better than CEL in all metrics. In TUH dataset, BUNDL uniformly improves FPR and latency for all models. Once

Table 4: Comparison of BUNDL and CEL using three state-of-the-art deep neural networks (DNN) across two datasets

DNN	Method	Simulated data			TUH data		
		FPR	Sensitivity	Latency	FPR	Sensitivity	Latency
DeepSOZ	BUNDL	2.6±0.8	0.90±0.06	30.1±21.7	2.3±1.1	0.74±0.08	7.6±10.9
	CEL	5.8±3.6	0.85±0.10	58.4±34.3	3.6±1.8	0.79±0.06	16.9±16.2
TGCN	BUNDL	2.63±0.85	0.89±0.04	191.6±37.6	2.6±2.5	0.75±0.09	30.1±29.7
	CEL	9.34±1.65	0.93±0.03	256.5±45.2	3.3±1.6	0.84±0.06	41.3±19.5
CNN	BUNDL	2.1±0.75	0.94±0.04	37.7±28.8	2.7±3.5	0.54±0.23	11.2±28.9
	CEL	4.6±1.36	0.94±0.05	45.6±21.4	3.0±2.2	0.66±0.13	15.5±20.1

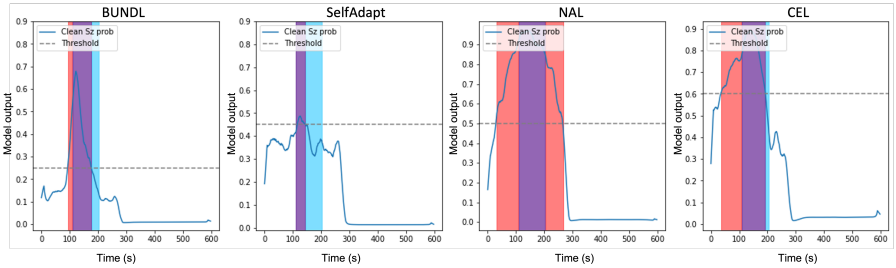

Fig. 3: Example seizure detection. Blue line is predicted clean seizure probability; dashed line is the detection threshold. Given seizure interval is in blue, predicted seizure interval is in red, and the intersection (correct prediction) is in purple.

again, we notice a slightly lower sensitivity for BUNDL, which can partially be attributed to computing the TUH evaluation metrics against the noisy clinician labels. We note that the sensitivity of BUNDL was highest in the simulated experiment, for which we had access to the "ground truth" clean labels.

Label Transition in TUH: We analyze the joint distribution $p(\hat{y}, y|x)$ by multiplying predicted $p(y|\hat{y}, x)$ with $p(\hat{y}|x)$ given by the test labels , to understand the behavior of noisy and predicted "clean" labels. Around onset region (i.e., 30-seconds before the annotated onset), the clinician labels are determined to be false positives by BUNDL with 0.21 probability (lower left quadrant), which reflects the ambiguous seizure start. During seizure, this probability rises to 0.383, due to increased uncertainty in EEG during seizure time points seen in Fig. 4(a) indicating that the clinician maybe mislabelling seizure onset. Effectively, BUNDL hedges towards baseline for highly uncertain intervals.

SOZ Localization: At seizure level, DeepSOZ trained using CEL has a localization accuracy of 0.601±0.176, which improves to 0.633±0.149 using BUNDL. Similarly at patient level, accounting for noisy labels using BUNDL increases the accuracy from 0.591±0.144 (CEL) to 0.620±0.111. This suggests that reducing ambiguities in detection helps to improve localization, as shown for one patient in Fig. 5. More examples are plotted in supplementary material.

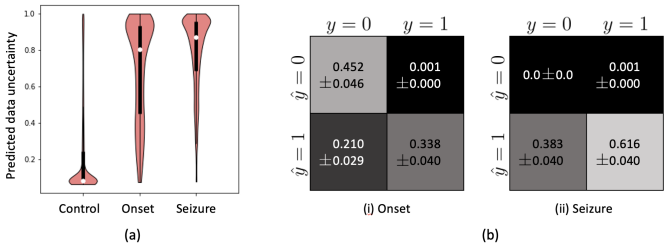

Fig. 4: (a) Predicted data uncertainty z_x in three time regions (b) Transition matrix - $p(\hat{y}, y|x)$.(i) Onset region and (ii) seizure region

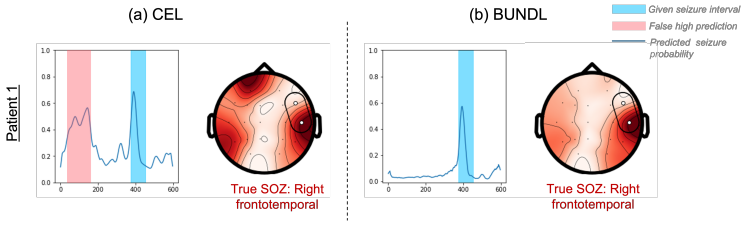

Fig. 5: Temporal seizure detection and seizure onset localization of two patients predicted by (a) DeepSOZ-BUNDL (b) DeepSOZ-CEL. Predicted SOZ is given by red heatmap. True SOZ is marked in black and mentioned underneath.

4 Conclusion

We have introduced, BUNDL, a novel statistical framework to mitigate the impact of noisy training labels. Specifically, BUNDL designs the transition between clean and noisy labels using a novel probabilistic model and facilitates robust training with a distributional loss. Using the backdrop of epileptic seizure detection, we demonstrated that training with BUNDL achieves improved performance as compared to state-of-art techniques in both simulated and real EEG data. Moreover, BUNDL can be used to train any existing deep learning model with minimal overhead. In an exploratory evaluation, we also demonstrate that addressing label noise in seizure detection enhances seizure onset zone localization, which is a critical step in epilepsy management. BUNDL is developed in a model-agnostic way and easily adaptable. Future work will explore its use in other medical imaging applications with various label noise types.

Acknowledgements

This work was supported by the National Institutes of Health R01 EB029977 (PI Caffo), the National Institutes of Health R01 HD108790 (PI Venkataraman), and the National Institutes of Health R21 CA263804 (PI Venkataraman).

References

1. Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U.R., et al.: A review of uncertainty quantification in deep learning: Techniques, applications and challenges. Information fusion **76**, 243–297 (2021)
2. Akbarian, B., et al.: A framework for seizure detection using effective connectivity, graph theory, and multi-level modular network. Biomedical Signal Processing and Control **59**, 101878 (2020)
3. Amin, U., Benbadis, S.R.: The role of eeg in the erroneous diagnosis of epilepsy. Journal of Clinical Neurophysiology **36**(4), 294–297 (2019)
4. Covert, I.C., et al.: Temporal graph convolutional networks for automatic seizure detection. In: Machine Learning for Healthcare Conference. pp. 160–180. PMLR (2019)
5. Craley, J., et al.: Automated inter-patient seizure detection using multichannel convolutional and recurrent neural networks. Biomedical signal processing and control **64**, 102360 (2021)
6. van Donselaar, C.A., et al.: Value of the electroencephalogram in adult patients with untreated idiopathic first seizures. Archives of neurology **49**(3), 231–237 (1992)
7. Goldberger, J., Ben-Reuven, E.: Training deep neural-networks using a noise adaptation layer. In: International conference on learning representations (2016)
8. Halford, J.J., Shiau, D., Desrochers, J., Kolls, B., Dean, B., Waters, C., Azar, N., Haas, K., Kutluay, E., Martz, G., et al.: Inter-rater agreement on identification of electrographic seizures and periodic discharges in icu eeg recordings. Clinical Neurophysiology **126**(9), 1661–1669 (2015)
9. He, J., et al.: Spatial–temporal seizure detection with graph attention network and bi-directional lstm architecture. Biomedical Signal Processing and Control **78**, 103908 (2022)
10. Huang, L., Zhang, C., Zhang, H.: Self-adaptive training: beyond empirical risk minimization. Advances in neural information processing systems **33**, 19365–19376 (2020)
11. Krol, L.R., Pawlitzki, J., Lotte, F., Gramann, K., Zander, T.O.: Sereega: Simulating event-related eeg activity. Journal of neuroscience methods **309**, 13–24 (2018)
12. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. Advances in neural information processing systems **30** (2017)
13. Li, C., et al.: Eeg-based seizure prediction via model uncertainty learning. IEEE Transactions on Neural Systems and Rehabilitation Engineering (2022)
14. M. Shama, D., Jing, J., Venkataraman, A.: Deepsoz: A robust deep model for joint temporal and spatial seizure onset localization from multichannel eeg data. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 184–194. Springer (2023)
15. Mnih, V., Hinton, G.E.: Learning to label aerial images from noisy data. In: Proceedings of the 29th International conference on machine learning (ICML-12). pp. 567–574 (2012)
16. Northcutt, C., Jiang, L., Chuang, I.: Confident learning: Estimating uncertainty in dataset labels. Journal of Artificial Intelligence Research **70**, 1373–1411 (2021)
17. Park, D., Choi, S., Kim, D., Song, H., Lee, J.G.: Robust data pruning under label noise via maximizing re-labeling accuracy. arXiv preprint arXiv:2311.01002 (2023)

18. Patrini, G., Rozza, A., Krishna Menon, A., Nock, R., Qu, L.: Making deep neural networks robust to label noise: A loss correction approach. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1944–1952 (2017)
19. Pedoeem, J., et al.: Tabs: Transformer based seizure detection. In: Biomedical Sensing and Analysis: Signal Processing in Medicine and Biology, pp. 133–160. Springer (2022)
20. Qiu, Y., Zhou, W., Yu, N., Du, P.: Denoising sparse autoencoder-based ictal eeg classification. IEEE Transactions on Neural Systems and Rehabilitation Engineering **26**(9), 1717–1726 (2018)
21. Reed, S., Lee, H., Anguelov, D., Szegedy, C., Erhan, D., Rabinovich, A.: Training deep neural networks on noisy labels with bootstrapping. arXiv preprint arXiv:1412.6596 (2014)
22. Shah, V., et al.: The temple university hospital seizure detection corpus. Frontiers in neuroinformatics **12**, 83 (2018), https://isip.piconepress.com/projects/tuh_eeg/html/downloads.shtml
23. Song, H., Kim, M., Park, D., Shin, Y., Lee, J.G.: Learning from noisy labels with deep neural networks: A survey. IEEE Transactions on Neural Networks and Learning Systems (2022)
24. Wang, H., et al.: Towards bayesian deep learning: A framework and some existing methods. IEEE Transactions on Knowledge and Data Engineering **28**(12), 3395–3408 (2016)
25. Xiao, T., Xia, T., Yang, Y., Huang, C., Wang, X.: Learning from massive noisy labeled data for image classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2691–2699 (2015)
26. Xue, C., Dou, Q., Shi, X., Chen, H., Heng, P.A.: Robust learning at noisy labeled medical images: Applied to skin lesion classification. In: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019). pp. 1280–1283. IEEE (2019)
27. Zhang, Y., et al.: Integration of 24 feature types to accurately detect and predict seizures using scalp eeg signals. Sensors **18**(5), 1372 (2018)
28. Zhang, Y., Niu, G., Sugiyama, M.: Learning noise transition matrix from only noisy labels via total variation regularization. In: International Conference on Machine Learning. pp. 12501–12512. PMLR (2021)