Automatic Report-Based Labelling of Clinical EEGs for Classifier Training

D. Western[1], T. Weber[1], R. Kandasamy[2], F. May, S. Taylor[3], Y. Zhu[1] and L. Canham[3]

1. Department of Engineering Design and Mathematics, UWE Bristol, Bristol, UK
2. National Hospital for Neurology and Neurosurgery, London, UK
3. North Bristol NHS Trust, Bristol, UK

david.western@uwe.ac.uk, timothy2.weber@live.uwe.ac.uk, rohan.kandasamy@nhs.net, drfelixmay@gmail.com, samantha.taylor@uhbw.nhs.uk, zhuyixuan1997@outlook.com, luke.canham@nbt.nhs.uk

Abstract— Machine learning classifiers for detection of abnormal clinical electroencephalography (EEG) signals have advanced signficantly in recent years, largely supported by the carefully curated Temple University Hospital Abnormal EEG Corpus (TUAB). Further progress towards clinically useful tools is likely to require larger volumes of data. In this study, we explore the viability and benefits of fully automated labelling of clinical EEG recordings based on the text in the clinical report, to efficiently exploit larger existing databases. We apply a machine learning classifier to the text reports in the Temple University Hospital EEG Corpus (TUEG) in order to label individual recordings. We show that training a deep convolutional neural network against the resulting dataset yields advantages in the resulting classification performance, namely increased area under the receiver operating characteristic curve and state-of-the-art specificity, albeit with a notable reduction in sensitivity. By demonstrating the viability of automatic report-based labelling, this paper opens the prospect of efficiently utilising the huge amount of historical EEG data in global medical archives to enhance the training of machine learning classifiers, either for enhanced general performance or bespoke training/evaluation for local populations.

I. Introduction

Electroencephalography (EEG) is a valuable investigation for a variety of conditions in medicine. Reporting the signals recorded from these investigations is time-consuming, and this limits the availability of the investigation. Efforts have been made recently to automate EEG analysis and reporting using deep learning [1]-[7] trained, in most cases, against the Temple University Hospital Abnormal EEG (TUAB) Corpus [8]. Determining if an EEG is normal, before delineating the nature of abnormalities, would be an important step towards having effective automated EEG screening.

For the training of these models, the availability of large volumes of labelled data is likely to be a limiting factor [9]. Data augmentation techniques have shown promise for enhancing deep learning for EEGs [10]. However, a large amount of authentic data has not been used to its full potential due to the effort required to robustly label the recordings. For many key tasks, such as classifying normal versus abnormal, a ground truth classification has already been performed by trained clinical professionals. In the Temple University Hospital EEG (TUEG) Corpus [11] and its derivative, TUAB, this information is stored alongside the EEG recording in the accompanying text report. Although this feature distinguishes the TUEG corpus from other research datasets, most clinical databases can be expected to hold some form of report to accompany EEGs. Hence, automated text-based classification of EEG reports as 'normal' or 'abnormal' would allow the labelling of large datasets to enhance training of models for EEG classification and other purposes.

Such automated labelling of historical datasets would inevitably yield differences from manually curated sets, such as TUAB. The nature and extent of how these differences will affect trained performance is unclear without empirical evidence. In this paper, we provide such evidence by evaluating a near-state-of-the-art model [1] trained against the carefully curated TUAB dataset in comparison with the same architecture trained against a larger derivative of TUEG, with data selected and labelled based on automated classification of report text (the reports written in semi-standardised natural language by health professionals).

II. Method

II-A. The TUAB and TUEG Datasets

As described by Lopes de Diego [8], although natural language processing (NLP) was initially used to label the TUAB dataset, all signals were manually reviewed to ensure accurate labelling based on inspection of the EEG recordings by a team of students. Where a single session contained multiple EDF (EEG recording) files, only a single file was taken to represent that session in the dataset. The dataset is split into training and evaluation sets, with 2,717 and 276 EDF files respectively, approximately balanced and demographically matched between normal and abnormal cases. Note that version 2.0.0 of TUAB is used throughout this study.

Our preliminary applications of text classification to the dataset exposed thirty-two cases in which the TUAB labelling did not agree with a human interpretation

of the report text. Such disagreements are expected, given that the TUAB labelling is based on inspection of the individual EDF file (not the full session) by a panel independent from the initial report author. The TUAB readme notes the existence of twenty-seven such cases. Given our focus on labelling from the reports, we relabelled the TUAB dataset such that the labels agreed with the reports. Our code repository includes a catalog of the disagreements and code to automatically relabel a copy of TUAB by moving files accordingly. This change affected only the training set, not the evaluation set.

In the larger TUEG dataset from which TUAB is derived, each session is accompanied by a report but may include multiple EDF files from across the session. Hence a session may include recordings with no evident abnormalities even when the report indicates that abnormality was identified within the session. In most cases, the reports do not contain any unambiguous indication of which files contain the evidence of abnormality.

II-B. Text Classifiers

We explored two different forms of text classifier applied to the reports, a simple rule-based (non-machine-learning) algorithm and natural language processing (NLP) based on a convolutional neural network.

II-B1. Rule-Based

Given that the majority of reports follow a similar format (factual report, describing objective features of the EEG, followed by a conclusion that summarisesd and interpretsed the findings for the clinician reading the report), and that there are well-defined degrees of abnormality associated with most EEG features (graphoelements), it is conceivable that a heuristic rule-based system could provide accurate classification.

We implemented a simple algorithm (Figure 1) that searches the first 80 characters in the conclusion section of the report for any of a set of terms indicating abnormality. If one is found, the report (and associated session) is labelled as ‘abnormal’. If not, an equivalent search is performed for the term ‘normal’. If found, the report is labelled as ‘normal’, otherwise it is labelled as ‘unknown’.

After refining this algorithm based on performance against the TUAB training set, it achieved 99.9% (2712/2716) accuracy using a set of just four key terms for abnormal: ‘abnormal’, ‘absence of normal’, ‘outside of the range of normal’, and ‘not normal’. Inspection revealed that the four errors were caused by absence of unambiguous key terms in the first part of the conclusion (two cases), a confounding phrase (“Normal EEG but abnormal EKG”), and a typo - “a normal” in place of “abnormal”, followed by a list of abnormalities observed. The algorithm was then evaluated against the TUAB test set and found to achieve 100% accuracy.

II-B2. Convolutional Neural Network

Although an extremely simple rule-based classifier was found to give good overall performance, the case of the typo demonstrated that such a system can be brittle. Furthermore, there is a high risk that it would not generalise well to datasets collected from other medical centers, which may use subtly different conventions in how reports are formatted and stored. Machine learning and NLP can potentially achieve a more robust classifier without substantial programming effort.

We implemented a CNN preceded by an embedding layer, as depicted in Figure 2. This conventional architecture is described more generally by Lauren *et al.* [12]. The embedding layer reduced the dimensionality of the data from a vocabulary of 6,514 words down to 64 concepts; each word is encoded as a vector in 64 dimensional space according to weightings that are trained along with the rest of the network. A single convolutional layer was used, with 64 kernels of size 5-by-64 (each spanning 5 word positions and the 64 embedding dimensions), a stride length of 2, ReLU activation functions, and global max pooling. The 64 outputs of the max pooling were reduced to two classes in a fully connected output layer with softmax activation.

The CNN was trained against the TUAB training set split 80%-20% for training and validation, then evaluated against the TUAB evaluation set, achieving 100% accuracy.

II-C. The AutoTUAB Dataset

We applied our text classifiers to the reports in the TUEG dataset to generate a larger, automatically labelled alternative to TUAB, which we name AutoTUAB for convenience. The classifiers disagreed with one another (i.e. the rule-based classifier reached a normal/abnormal verdict and it was different from that of the CNN) in 546 out of 26,387 reports. A preliminary inspection of these cases of disagreement indicated that the probability of abnormality returned by the CNN was less conclusive than in cases of agreement. As depicted in Figure 3, the CNN returned probabilities in the ranges <0.01 and >0.99 for 83% of the reports classified. Hence it was decided to use the CNN alone with these thresholds to identify confident classifications and exclude the less confident ones.

In keeping with the fact that TUAB contains no overlap between subjects in the training and evaluation sets, recordings from subjects included in TUAB’s evaluation set were excluded from the AutoTUAB training set. For consistency of comparisons with other studies, we used the TUAB evaluation set as the test set for all trained models.

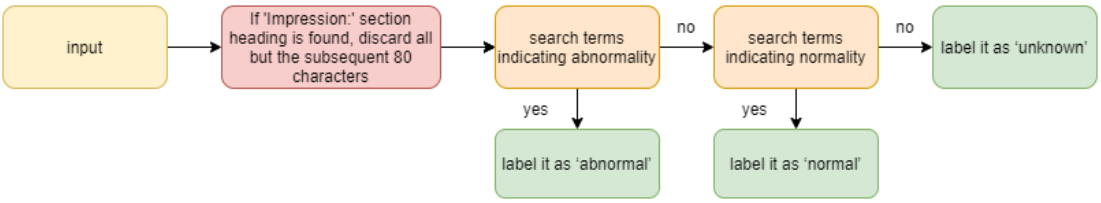

Figure 1. A flowchart depicting the rule-based text classifier

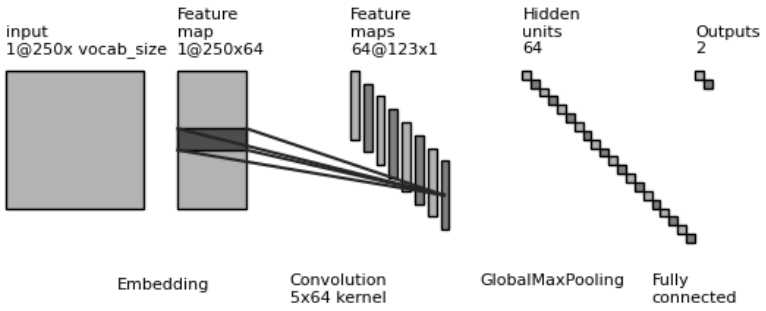

Figure 2. Architecture of the CNN text classifier, depicted as the transforming shape of the data representation as it moves through the network (left to right)

The TUEG dataset contains substantially more abnormal than normal recordings. To avoid biasing the model, a balanced variant of AutoTUAB was created by discarding abnormal recordings to achieve the same number of normal and abnormal examples. To maximise the number of subjects retained, files were selected at random while avoiding repeat selections from any given subject until all subjects had contributed at least one file. This process resulted in 8701 abnormal files (reduced from 27,815) from 8,261 subjects and 8,701 normal files.

II-D. EEG Classifier

For simplicity, our study focussed on a single previously developed model for classifying EEGs as normal/abnormal from the signals themselves. The BD-Deep4 CNN [1] provides near-state-of-the-art performance (see Table 1) with a broadly conventional architecture and open source implementation. Schirrmeister *et al*'s study was among the first to apply CNN's to clinical EEG classification. It included the important observation that reducing the considered recording duration from 20 minutes to 1 minute (the second minute of the recording, because the artefact-prone first minute was always skipped) incurred only a small loss in accuracy, from 85.4 percent to approximately 82 percent.

II-D1. Architecture

The original model architecture was unaltered throughout our study. For a full description, see [1]. In brief, it consists of four convolution blocks and two softmax output layers.

II-D2. Training and Evaluation

Training and evaluation were conducted using the original code provided by Schirrmeister *et al.* [1] with minimal changes as necessary to implement our changes to the training data without requiring further adaptations to the TUAB/TUEG datasets.

To mitigate the increased computational expense of a larger training dataset, we used only the second minute of each recording (the first minute is discarded due to increased prevalence of artefacts). Recordings shorter than 2 minutes were excluded from the training data - none were present in the evaluation data.

Table 1. Summary of state-of-the-art performance metrics for different models applied to abnormal EEG classification

Model	Accuracy	Sensitivity	Specificity
BD-Deep4[1]	85.4 %	75.1 %	94.1 %
1D-CNN (T5-O1 channel)[2]	79.3 %	71.4 %	86.0 %
1D-CNN (F4-C4 channel)[2]	74.4 %	55.6 %	90.7 %
CNN-MLP[8]	78.8 %	75.4 %	81.9 %
KNN[8]	58.2 %	66.0 %	50.5 %
VGG-16 + SVM[13]	86.6 %	77.8 %	94.0 %
AlexNet + SVM[13]	87.3 %	78.6 %	94.7 %
3 x AlexNet + MLP[14]	89.1 %	80.2 %	96.7 %
GMM-HMM-SdA[8]	75.4 %	90.0 %	62.3 %
HMM[3]	83.8 %	86.8 %	82.3 %
HMM-SdA[3]	90.1 %	78.9 %	95.6 %
HMM-SdA-SLM[3]	93.5 %	90.1 %	95.1 %

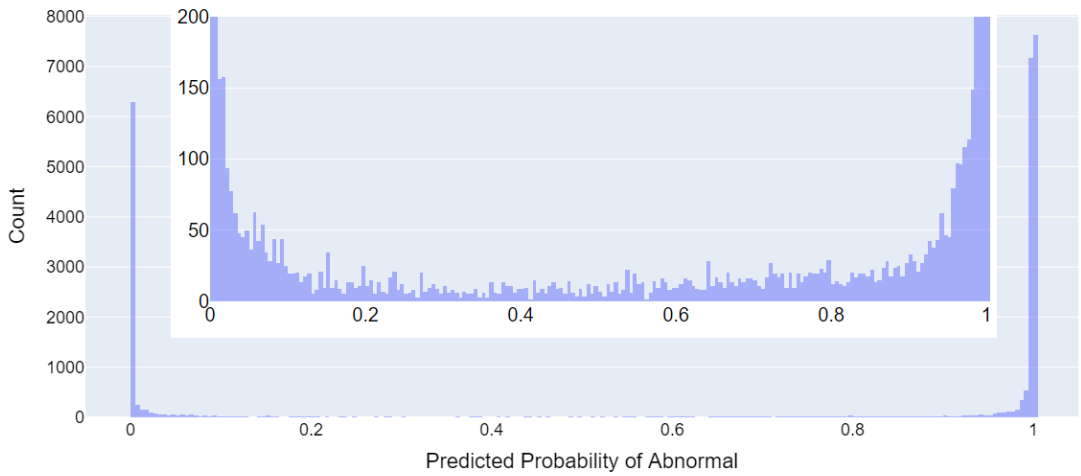

Figure 3. Histogram of 'probability of abnormal' (p_{ab}) values attributed to reports in TUEG by the CNN classifier. Inset is the same histogram with the vertical scale adjusted to clarify the smaller bars in the central portion.

In keeping with the original implementation of BD-Deep4 [1], up to 20 minutes was used from each recording in the evaluation dataset, and recordings of longer than 35 minutes were excluded from the training data, presumably to manage memory constraints.

All EEG classifier models were assessed based on accuracy, sensitivity, specificity, and the area under the receiver operating characteristic curve (AUC).

III. Results

III-A. TUAB Relabelled

As noted in Section II-A, we relabelled thirty-two recordings (1%) from TUAB to ensure consistency with the report text. Although the selection of recordings was informed by inspection of the signals, the labelling of this variant of TUAB was, in effect, solely based on the report text. To determine the effect of this change on baseline performance, we trained the BD-Deep4 architecture against both the original and relabelled TUAB variants, using up to 20 minutes per recording. As shown in Table 2, the relabelling resulted, on average, in a small degradation of overall accuracy and sensitivity and a small increase in specificity. However, there was considerable variation between runs for the relabelled data, of which the best accuracy and sensitivity exceeded that of the original training data. In summary, the results in Table 2 confirm that relabelling the original TUAB dataset based on the report text does not substantially affect performance, presumably because the report-based labels agreed with the originals in 99 percent of cases.

III-B. AutoTUAB

Table 3 shows that training against the AutoTUAB dataset resulted in a decrease in overall accuracy and sensitivity but an increase in specificity.

IV. Discussion

The results in Tables 2 and 3 suggest that labelling TUEG EEG recordings based on session reports (as in AutoTUAB and the relabelled TUAB) rather than the observed signal content reduces the sensitivity of the trained model but increases specificity. Report-based labelling presents three fundamental differences from the original TUAB in terms of the nature of the training data:

- Exposure to recordings labelled 'abnormal' but without obvious abnormalities in the signal, perhaps confounding its learned representation of this class (reduced sensitivity).

Table 2. Performance metrics averaged over 10 runs. The larger value for each column is shown in bold. 20 minutes was used from each recording.

Training Data	Accuracy	Sensitivity	Specificity	AUC
Original TUAB	**85.9 %**	**77.0 %**	93.3 %	**0.917**
Relabelled TUAB	85.4 %	75.2 %	**94.0 %**	0.916

Table 3. Performance metrics averaged over 5 runs. The larger value for each column is shown in bold. 1 minute was used from each recording.

Training Data	Accuracy	Sensitivity	Specificity	AUC
Original TUAB	**81.8 %**	**71.0 %**	90.9 %	0.909
AutoTUAB	77.9 %	53.7 %	**98.3 %**	**0.912**

- Non-exposure to recordings labelled 'normal' from sessions in which abnormalities were observed, which might otherwise disrupt the model's ability to learn useful features not taken into account by human interpreters.

- Sub-optimal (random) selection of the signals from within each session, degrading the overall quality of data used in training.

A central hypothesis in our work was that this degradation in overall data quality could be mitigated by an increase in the volume of data considered. The results do not provide a straightforward confirmation or rejection of this hypothesis. Overall accuracy was lower for AutoTUAB, but the area under the receiver operating characteristic curve was greater. The improvement in specificity is notable in that 98.3% arguably represents a clinically useful value for some applications, such as 'ruling in' suspected neurophysiological pathology.

To our knowledge (see Table 1), 98.3% is the highest value of specificity reported for a model tested against TUAB. However, the result should be interpreted cautiously given the low sensitivity of the AutoTUAB-trained model; the model is apparently biased towards returning a prediction of 'normal', despite being trained on a balanced dataset. Nonetheless, the improvement in AUC indicates that the improvement in specificity cannot solely be explained as bias, since AUC is independent of the decision threshold.

The scope of this investigation was limited by practical constraints, but the noted improvements warrant further work. Several important parameters have yet to be explored.

- **Architecture:** We have only used a single model architecture, whereas it is conceivable that the effects of training against AutoTUAB may vary between other state-of-the-art architectures.

- **Duration:** Due to computational resource constraints, AutoTUAB-training was only performed using one minute from each recording; in further work we will investigate whether more pronounced benefits are achieved when more of the signal is available to learn from.

- **Selection of recording section:** The second minute of each recording was used for training. It is possible that performance could be improved by introducing an algorithm to select a recording section based on statistical analyses. However, we do not believe that this consideration influences the discrepancies between the two rows of Table 3, because both training sets used the same approach in this regard, yet the changes from the baseline of Table 2 were very different. We believe that a more influential factor in the discrepancy was the fact that, unlike the original TUAB, the recording selected from each session for AutoTUAB was not necessarily one that substantially influenced the choice of label (i.e. the clinical conclusion in the report).

- **Test set:** All models in our investigation were evaluated against the TUAB evaluation set. This arguably favours training with the original TUAB training set, which would have greater similarity to the test set than AutoTUAB would.

- **Per-session classification:** We noted that 'normal' recordings within 'abnormal' EEG sessions are potentially confounding to machine learning. In our report-based labelling, all such recordings are classed as 'abnormal', granting the opportunity for the model to learn useful features that go undetected by human interpreters. Since the conclusions of an EEG report apply to the session rather than specific recordings, it is natural to consider a paradigm shift in EEG classification, such that the classification is based collectively on all recordings within a session rather than on individual files. This would require significant adaptation to established architectures and/or training algorithms, but could be key in allowing machine learning to exceed human performance in this domain.

V. Conclusion

Automatic report-based labelling of clinical EEG sessions provides distinct advantages in training classifiers to distinguish between normal and abnormal recordings. We demonstrated an improvement in AUC compared with training against the original TUAB training set. We also achieved state-of-the-art specificity against the TUAB evaluation set, albeit with notably low sensitivity. Carefully curated datasets such as TUAB will always play an important role in the training and evaluation of EEG classifiers. However, the viability of automatic report-based labelling suggests that the huge amount of historical EEG data stored in archives of medical centers around the world could be relatively efficiently deployed towards training of machine learning models. Benefits may include enhanced performance in the general population as well as bespoke training/evaluation for local populations.

Code and Supplementary Data

For the sake of reproducibility and enabling others to work with the AutoTUAB dataset, our code repository includes the following:

- The code used to define and train the text classifier.
- A catalog of the resulting 'probability of abnormal' values and associated report file paths within TUEG.
- Code demonstrating their use to apply labels to TUEG and train BD-Deep4 (forked from [1] and adapted).

Available at https://github.com/DWonGH/autotuab.

ACKNOWLEDGEMENTS

D. Western's contribution was supported by a UWE Bristol Wallscourt Fellowship and Vice Chancellor's Early Career Researcher Award. T. Weber's contribution was supported by a research internship funded by the UWE Bristol Faculty of Environment and Technology.

REFERENCES

[1] R. T. Schirrmeister, L. Gemein, K. Eggensperger, F. Hutter, and T. Ball, "Deep Learning with Convolutional Neural Networks for Decoding and Visualization of EEG Pathology," *arXiv e-prints*, p. arXiv:1708.08012, Aug. 2017. (available at: *https://ui.adsabs.harvard.edu/abs/2017arXiv170808012T*).

[2] Ö. Yıldırım, U. B. Baloglu, and U. R. Acharya, "A Deep Convolutional Neural Network Model for Automated Identification of Abnormal EEG Signals," *Neural Computing and Applications*, Nov. 2018. (available at: *https://doi.org/10.1007/s00521-018-3889-z*).

[3] M. Golmohammadi, A. H. Harati Nejad Torbati, S. Lopez de Diego, I. Obeid, and J. Picone, "Automatic Analysis of EEGs Using Big Data and Hybrid Deep Learning Architectures," *Frontiers in Human Neuroscience*, vol. 13, 2019. (available at: *https://www.frontiersin.org/articles/10.3389/fnhum.2019.00076/full*).

[4] Y. Roy, H. Banville, I. Albuquerque, A. Gramfort, T. H. Falk, and J. Faubert, "Deep Learning-Based Electroencephalography Analysis: A Systematic Review," *Journal of Neural Engineering*, vol. 16, no. 5, p. 051001, Aug. 2019. (available at: *https://doi.org/10.1088/1741-2552/ab260c*).

[5] S. Roy, I. Kiral-Kornek, and S. Harrer, "ChronoNet: A Deep Recurrent Neural Network for Abnormal EEG Identification," *Artificial Intelligence in Medicine*, ser. Lecture Notes in Computer Science, D. Riaño, S. Wilk, and A. ten Teije, Eds. Cham: Springer International Publishing, 2019, pp. 47–56.

[6] S. Biswal, C. Xiao, L. Glass, B. Westover, and J. Sun, "Clinical Report Auto-completion," *Proceedings of The Web Conference 2020*, ser. WWW '20. Taipei, Taiwan: Association for Computing Machinery, Apr. 2020, pp. 541–550. (available at: *https://doi.org/10.1145/3366423.3380137*).

[7] J. Thiagarajan, D. Rajan, S. Katoch, and A. Spanias, "DDxNet: A Deep Learning Model for Automatic Interpretation of Electronic Health Records, Electrocardiograms and Electroencephalograms," *Scientific Reports*, vol. 10, no. 1, p. 16428, Oct. 2020. (available at: *https://www.nature.com/articles/s41598-020-73126-9*).

[8] S. Lopez de Diego, "Automated Interpretation of Abnormal Adult Electroencephalograms," Masters, Temple University, 2017. (available at: *https://scholarshare.temple.edu/handle/20.500.12613/1767*).

[9] L. Gemein, R. Schirrmeister, P. Chrabąszcz, D. Wilson, J. Boedecker, A. Schulze-Bonhage, F. Hutter, and T. Ball, "Machine-Learning-Based Diagnostics of EEG Pathology," *NeuroImage*, vol. 220, p. 117021, Oct. 2020. (available at: *https://www.sciencedirect.com/science/article/pii/S1053811920305073*).

[10] E. Lashgari, D. Liang, and U. Maoz, "Data Augmentation for Deep-Learning-Based Electroencephalography," *Journal of Neuroscience Methods*, vol. 346, p. 108885, Dec. 2020. (available at: *https://www.sciencedirect.com/science/article/pii/S0165027020303083*).

[11] I. Obeid and J. Picone, "The Temple University Hospital EEG Data Corpus," *Frontiers in Neuroscience*, vol. 10, 2016. (available at: *https://www.frontiersin.org/articles/10.3389/fnins.2016.00196/full*).

[12] P. Lauren, G. Qu, and P. Watta, "Convolutional Neural Network for Clinical Narrative Categorization," *2017 IEEE International Conference on Big Data (Big Data)*, 2017, pp. 2001–2008.

[13] S. Amin, M. S. Hossain, G. Muhammad, M. Alhussein, and A. Rahman, "Cognitive Smart Healthcare for Pathology Detection and Monitoring," *IEEE Access*, vol. PP, pp. 1–1, 01 2019.

[14] M. Alhussein, G. Muhammad, and M. S. Hossain, "EEG Pathology Detection Based on Deep Learning," *IEEE Access*, vol. 7, pp. 27 781–27 788, 2019.